Universidad Nacional Autónoma de México

Centro de Física Aplicada y Tecnología Avanzada

Análisis de deleciones en el Genoma Humano

QUE PARA OBTENER EL TÍTULO DE:

Licenciado en Tecnología

PRESENTA:

Edgar Iván Chávez Aparicio

TUTOR:
Dra. Maribel Hernández Rosáles
COTUTOR:
Dr. Alfredo Varela Echavarría

Santiago de Querétaro, Querétaro, 2020

A todas las personas que directamente o indirectamente me ayudaron a llegar a este punto en mi vida:

- *A mi mamá (María Eugenia Aparicio Alegría)*
- *A mi papá (Esteban Chávez Cano)*
- *A toda la familia Aparicio*
- *A Todo el clan de los Chávez*
- *A Andrés García García (Alias "El mano Loco") y su novia Fernanda Espejo ("La chica Loca")*
- *A Lucero Mescli Hernández*
- *A Jaime Luis De Santiago Cáravez*
- *A Marcos Emmannuel Gonzáles Laffite*
- *A Alitzel López Sánchez*
- *A Ezequiel Gonzáles Cruz*
- *A Héctor Colt Celaya*
- *A Julio Alberto Rodrigo Franco Guerrero*

Así como a Vito's Pizza y Tacos Don Agus por estar ahí cuando los necesitaba. Además quisiera dedicar esta tesis a memoria de Stephen Hillenburg (1961 - 2018), biólogo marino y creador de la caricatura "Bob Esponja".

Reconocimientos

Este trabajo se llevó a cabo con apoyo del Laboratorio Nacional de Visualización Científica Avanzada (LAVIS) y de Luis Alberto Aguilar Bautista y Alejandro de León Cuevas.

Se agradece la asistencia técnica del Ing. Carlos Sair Flores Bautista.

Resumen

En los organismos pluricelulares existen dos tipos de células, las germinales y las somáticas; las primeras corresponden a las que tienen la capacidad de ser fecundadas para formar un nuevo organismo independiente y las segundas a las células que componen el cuerpo del organismo. En ambos casos pueden existir alteraciones en la información genética de las células, debido a la duplicación celular o el ambiente, a lo cual se le llama mutaciones. En este proyecto se realiza un nuevo método para entender las mutaciones que ocurren en las células somáticas. La importancia de esto radica en que dichos eventos están relacionados con los mecanismos que se han propuesto en cuanto al envejecimiento y el cáncer (6, 32, 59, 60).

Durante el desarrollo de organismos pluricelulares, las células pasan por modificaciones que les confieren especificidad. Sin embargo, aún es necesario seguir indagando sobre la naturaleza y los mecanismos que provocan mutaciones en el ADN a lo largo del desarrollo. Entender cuáles mutaciones son consecuencia de un mecanismo biológico permitirá entender mejor los procesos de envejecimiento y carcinogénesis. El ojetivo de este trabajo es identificar deleciones somáticas similares a las reportadas en (20, 31, 32, 34, 46, 62) a partir de datos de secuenciación completa en seres humanos. Esta información será útil para comparar el tipo de eventos que se llevan a cabo de forma natural en distintos tejidos de un adulto. No obstante, estos varían en cuanto a longitud, comprendiendo desde una base hasta cromosomas enteros, así como en especificidad de especie o tejido.

Hasta el momento no se ha encontrado la función que tienen dichos eventos, pero en ciertos casos las observaciones son recurrentes, como en el caso de los nematodos con estilo de vida parásita (46), lo que apoya la posibilidad de que estos eventos sean parte del proceso natural de desarrollo. Delimitamos este trabajo a células humanas. En lo particular buscamos los eventos que sean compartidos entre distintos individuos sanos. Para cumplir el objetivo se desarrollaron herramientas computacionales que pueden procesar grandes cantidades de datos de secuenciación masiva.

Los resultados muestran deleciones de origen somático compartidas entre distintos experimentos de secuenciación. La naturaleza de ellas es variada en cuanto a ubicación y longitud. Empero, existen falsos positivos, es decir, deleciones detectadas como somáticas que en realidad son heredadas desde líneas germinales, a consecuencia de

errores de alineamiento o deleciones dispersas en regiones de alta variabilidad que no pueden ser comprobadas experimentalmente.

Esta es una aproximación que permite corrobrar la funcionalidad de la herramienta para comprobar la existencia de deleciones, incluidas aquellas de baja frecuencia, como parte del desarrollo normal en el ser humano. Se utilizaron cuatro experimentos, dos provenientes de tejido de seno, y dos provenientes de sangre. En este trabajo que consideraron deleciones recurrentes aquellas que sobrelapaban en por lo menos un 50 % entre todos los individuos. Como trabajo a futuro, es posible extender este método para descubrir deleciones cuyo traslape sea distinto al 50 %.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En el ser humano existen distintas variantes genéticas que se presentan a lo largo de la vida y el interés de esta investigación recae en las deleciones somáticas en una población de células sanas que ocurren en el organismo completamente formado, y que son recurrentes entre distintos individuos. Para ello, se desarrollaron y aplicaron herramientas computacionales necesarias para hacer la detección de deleciones en ADN en células somáticas.

Por el momento, los programas para la detección de deleciones somáticas se enfocan en regiones con genes (26, 47), están enfocados a estudiar enfermedades, requieren métodos experimentales específicos (12, 23) o cuantifican magnitudes de mutaciones (43). Sin emabrgo, es necesario desarrollar herramientas que ayuden a vislumbrar la naturaleza de las deleciones somáticas a partir de los experimentos de secuenciación, las cuales estén presentes en baja frecuencia o sean descartadas por otros programas de detección de variantes, de manera similar a lo que se reporta en estudios llevados a cabo sobre lineas celulares modificicadas (49, 54); esto puede proporcionar información sobre padecimientos relacionados con el envejecimiento y ciertas enfermedades como el cáncer.

Las tecnologías actuales de secuenciación masiva permiten conocer los genomas nucleares contenidos en las células, ya que son capaces de leer una gran cantidad de cadenas de ADN en un tiempo reducido. Por lo tanto, es posible buscar variantes genéticas en células somáticas del tejido humano.

En este trabajo se analizaron cuatro experimentos de secuenciación en aras de obtener deleciones de origen somático que estuvieran compartidas entre todos los individuos. Estas corresponden a dos experimentos en sangre; “SRX237626” y “SRX1660320” con tamaños de 86.1 y 249.4 gigabases respectivamente; y dos en seno; “SRX257065” y “SRX257065” con 426 y 213.7 gigabases respectivamente.

El producto de este trabajo es un conjunto de herramientas capaces de realizar un análisis bioinformático integral que lleva a cabo desde la descarga de datos de secuenciación hasta la comparación de deleciones entre distintos experimentos, pasando por la detección de deleciones individuales, el filtro de regiones repetitivas, deleciones que no sean somáticas, entre otros análisis. El software fue implementado en el servidor

DNA del Laboratorio Nacional de Visualización Científica Avanzada (LAVIS) donde se analizaron los datos.

1.1. Fundamentación y justificación

La biología de desarrollo es una rama de la biología que se dedica a entender la organización que conforma a los organismos vivos (61). Este desarrollo comprende desde la concepción hasta su proceso de muerte, ocasionada por los procesos de senescencia. Los procesos y mecanismos específicos son sujetos de estudio en la actualidad, por lo que se van realizando nuevos descubrimientos conforme el área de investigación crece.

Para el estudio de un organismo pluricelular se pueden clasificar las células en dos grupos distintos (61). Al primero se le conoce como células germinales, las cuales son capaces de formar otro organismo. Al segundo se le conoce como células somáticas que conforman los tejidos especializados del organismo.

Estas últimas están sujetas a diversas modificaciones que pueden cambiar el nivel de expresión de los genes. Dichas modificaciones confieren especificidad a la célula, así como algunas facultades dentro de un tejido u órgano. La especificación se va realizando conforme las células del óvulo fecundado se van dividiendo en las primeras etapas de desarrollo dando origen a células somáticas (61).

1.1.1. Mutaciones somáticas

Las células pueden sufrir alteraciones que tengan como consecuencia modificaciones en su ADN. Estos cambios pueden ser deleciones que generan pérdidas de un segmento de ADN; puede haber inserciones, que introducen bases o sustituciones, mismas que cambian el tipo de base en una ubicación. También pueden existir cambios mayores a una base, tales como las reversiones, que invierten la secuencia en una región; duplicaciones, que repiten fragmentos de ADN; asimismo, traslocaciones, que cambian la ubicación de un segmento de ADN, así como deleciones de mayor extensión (6).

El estudio de procesos de mutación en células somáticas es de relevancia para comprender cuáles son los cambios que se dan en nuestras células como parte del crecimiento y el envejecimiento. Se requiere realizar nuevas investigaciones para encontrar y comprender patrones en las modificaciones de células somáticas. Para ello es necesario identificar aquellas que suceden como parte del desarrollo natural.

En esta tesis el enfoque recae en las mutaciones deletéreas de cualquier longitud; estas modificaciones son aquellas que remueven un conjunto de bases consecutivas de ADN, las cuales pueden presentarse de forma aleatoria, causada por errores durante la duplicación de ADN, o bien en ubicaciones o fases del desarrollo específicas. A continuación son presentados distintos eventos deletéros de ADN a escala somática que ocurren en distintos tipos de células.

Mutaciones durante el desarrollo de distintas especies

Ciertas especies de nematodos presentan deleciones de ADN de regiones extensas en las células que pasan a ser somáticas durante las primeras divisiones celulares (20, 34, 46). Este tipo de evento también ocurre en ciertos crustáceos (46), protozoarios ciliados (46), peces agnatos (34) e incluso en algunos marsupiales (34). En el caso de las moscas, hay especies que pasan por procesos que eliminan cromosomas enteros para dar paso a células somáticas a partir de células germinales (21).

Neuronas

Los ratones (31) y la salamandra *Cynops Pyrrhogaster* (62) pasan por procesos deletéreos durante etapas tempranas del desarrollo. Estos cambios afectan a células que dan paso a sus neuronas.

Sistema inmune

En vertebrados existen rearreglos genéticos en las células B y T del sistema inmune, las cuales dan origen a una gran variedad de posibles anticuerpos para distintos antígenos (41). Estos cambios en la genética en las células B y T son un mecanismo conservado que dotan a las células de una función específica. Dichas modificaciones no son aleatorias, y es posible que existan mecanismos similares en otros tejidos del humano.

Envejecimiento

El envejecimiento es un proceso natural en los organismos pluricelulares (13). Uno de los modelos actuales indica que este es consecuencia de la pérdida de las funciones celulares de manera progresiva. Las modificaciones en las funciones celulares parecen deberse a los cambios de expresión genética que se dan tras alterar regiones del ADN (6, 60). A pesar de esto, la explicación completa es desconocida aún y permite especular posibles relaciones con las deleciones somáticas.

Cáncer

Además del envejecimiento, el cáncer parece ser efecto de la acumulación de mutaciones somáticas (32). De manera específica, aquellas mutaciones relacionadas con los oncogenes y los genes supresores de tumores; las demás mutaciones son llamadas pasajeras. La respuesta de cómo se crean las primeras células carcinógenas, o de cómo evitar su creación, puede estar explicada por mecanismos que inducen mutaciones.

1.1.2. Detección de variaciones somáticas

Las variaciones genéticas se refieren a las diferencias entre dos o más secuencias de ADN comparables. Es posible comparar secuencias si hay alguna forma de indicar que tienen una misma ubicación, referencia o función semejante. Esto depende de la escala a la cual sea realizada la comparación. Entre especies es posible comparar regiones homólogas, las cuales tienen funciones o provienen de un gen ancestro común. También es posible comparar dos regiones que se encuentren dentro de un mismo intervalo en el genoma de una especie.

Una variación genética es una forma más general de una mutación. En particular, las variaciones somáticas son un subconjunto de variaciones genéticas que se presen-

tan debido a las divisiones celulares que se llevan a cabo durante el desarrollo de un organismo.

Es posible identificar las variantes genéticas que están presentes en distintas células realizando una comparación base por base de las cadenas de ADN. Esto es útil para identificar variaciones somáticas.

Esto es gracias al avance de las nuevas tecnologías de secuenciación. A partir de esta información es posible identificar diferencias y similitudes entre muestras biológicas.

Actualmente existen varios programas que realizan la identificación de variantes genéticas en humanos. Es decir, que revisan cuáles son los cambios que tiene una muestra en su ADN con respecto a otras células de un individuo. Dichas variantes se encuentran en una proporción del total de células, por lo que la principal aplicación de estas herramientas es la detección de variantes genómicas en células cancerígenas.

Cuando las células con cáncer son analizadas, unas herramientas consideran la presencia de poliploidía para llevar a cabo la identificación. Sin embargo, estámos enfocados en analizar células sanas que sean diploides.

Por tanto, buscamos programas que encuentren variaciones somáticas sin suponer poliploidía. Tal es el caso de `SomatiCA` (8) y de `Control-FREEC` (3). Por el otro lado, existen otros programas que determinan las deleciones presentes en los alelos de células diploides. Un ejemplo de esto es `POD` (2), el cual tiene la desventaja de requerir información genómica de los padres del individuo, lo cual restringe los estudios a familias en vez de individuos.

Sin embargo, el interés de esta investigación recae en las deleciones somáticas en una población de células sanas que ocurren en el organismo completamente formado. Para esto, es necesario considerar el subconjunto de deleciones somáticas que ocurren en los tejidos de un individuo.

Igualmente, existen herramientas para la detección de deleciones, tales como `GATK`, `SoloDel` (33), `LUMPY` (36), `DELLY` (52), `Sprites` (63), `MosaicHunter` (25), `VarScan2` (35), entre otros. Sin embargo, unas herramientas requieren más de un tejido de muestra con el fin de llevar a cabo comparaciones, en contraste, nosotros necesitamos llevar a cabo en análisis con uno solo. Por el otro lado, otras se enfocan ya sea en deleciones de una sola base, o variaciones estructurales, las cuáles son muy grandes, de igual forma existen programas como `LUMPY` (36) que integran información de ambas aproximaciones. Sin embargo, nosotros nos enfocamos en deleciones de cualquier tamaño posible inferidas a partir de un enfoque similar a analizar base por base.

1.2. Hipótesis

Si bien una función específica a la presencia de deleciones no se ha logrado comprobar, esto permite preguntarse si es posible que existan eventos compartidos entre los tejidos somáticos del humano. En lo particular, aquellos cambios que no sean consecuencia de algún padecimiento.

En esta tesis se propone la existencia de deleciones somáticas como parte del desarrollo normal del ser humano completamente formado. De ser así, entonces estas estarán presentes en más de un experimento de secuenciación en personas sanas.

1.3. Planteamiento del problema

En este trabajo se buscan todas las deleciones que se puedan catalogar como somáticas dentro de distintos experimentos de secuenciación en el ser humano. Se pretende buscar aquellas modificaciones de ADN que no sean consecuencia de patologías o la metodología experimental. Esto tiene como fin que las deleciones identificadas como recurrentes entre distintos individuos, sean solo aquellas que estén relacionadas con el desarrollo a lo largo de la vida del individuo. Las modificaciones compartidas entre individuos podrían ser corroboradas experimentalmente en trabajos a futuro. Para realizar la corroboración experimental, es necesario evitar las regiones de ADN repetido, ya que los procesos experimentales a disposición en el grupo de trabajo son de baja fidelidad en dichas regiones.

Bajo estos requerimientos se llevó a cabo el desarrollo de las herramientas bioinformáticas necesarias para realizar el análisis completo. Estas herramientas deben contar con la suficiente flexibilidad para poder realizar análisis similares en otros trabajos. Además, es necesario que el proceso esté automatizado para ser utilizado en otros proyectos.

1.4. Objetivo

Desarrollar las herramientas computacionales necesarias para poder realizar la detección de deleciones de origen somático, presentes en el desarrollo normal del humano, a partir de datos de secuenciación masiva. Este trabajo se enfoca en buscar las deleciones que cumplan con un mínimo de restricciones, esto con el fin de ampliar la cantidad posibles deleciones somáticas.

1.4.1. Objetivos intermedios

En este trabajo se desarrollaron herrameintas bioinformáticas para llevar a cabo las siguientes tareas:

- Analizar, curar y mapear al genoma de referencia de las lecturas obtenidas de la secuenciación.
- Detección de deleciones en lecturas alineadas.
- Filtros de deleciones en zonas repetidas, así como aquellas de origen germinal.

- Identificación de deleciones compartidas entre experimentos.
- Análisis de las características de las deleciones somáticas encontradas.

1.5. Metodología

Este trabajo requirió una búsqueda de datos de secuenciación en bases de datos públicas, los cuales se utilizaron para corroborar la presencia de deleciones. Se utilizaron datos crudos de secuenciación en células de sangre y seno, provenientes de cuatro individuos distintos, como primera aproximación. La información fue obtenida de la base de datos de "SRA". Posteriormente, fueron obtenidas las ubicaciones de las deleciones somáticas detectadas en el genoma humano, las cuales no estuvieran próximas a regiones repetitivas del genoma. Las ubicaciones resultantes fueron corroboradas entre todos los experimentos analizados. Adicionalmente, en los registros fueron comparadas las ubicaciones resultantes con los genes asociados con cáncer, los cuales fueron obtenidos de la base de datos "Genomic Data Commons" (40).

1.6. Contribuciones

En este trabajo se desarrollaron las herramientas bioinformáticas a fin de analizar datos crudos de secuenciación, con los cuales se pueden identificar las deleciones somáticas recurrentes entre distintos experimentos. Con estas herramientas se logró encontrar evidencia de la existencia de las deleciones mencionadas, las cuales ocupaban las mismas coordenadas en cuatro individuos distintos.

Con el conjunto de herramientas, el usuario será capaz de realizar el análisis de forma automática, lo que presenta una mayor facilidad de operación a usuarios futuros. De igual manera, se podrán ajustar los parámetros para emprender distintos análisis de eventos somáticos. Además, es posible que sean anexadas utilidades para refinar resultados o filtrar otro tipo de artificio.

La implementación se llevó a cabo en el servidor llamado "DNA" en el LAVIS. Estas herramientas pueden ser utilizadas por cualquier otro usuario, las cuales se encuentran en un repositorio en línea (`https://github.com/Bloodfield/Plasticidad_INB`).

1.7. Estructura de la tesis

Este proyecto surgió bajo la necesidad de responder preguntas dentro del área de la biología, las cuales son resueltas desde la perspectiva bioinformática. Por ello se presenta primero un panorama de las ciencias genómicas y la bioinformática. Como parte de ello, se exponen las tecnologías actuales de secuenciación masiva y se explican los fundamentos para realizar los análisis de este proyecto.

Posteriormente, se explican los fundamentos biológicos que dan sustento a la hipótesis, se presentan distintas células dentro de los mamíferos y organismos que pasan por mutaciones somáticas programadas. En ciertos casos la función de dicho proceso es desconocida, empero, permite preguntarse si existen en el ser humano eventos similares, es decir, eventos de deleción como parte del desarrollo natural.

Una vez presentada la motivación biológica, se pasa a explicar la detección de deleciones somáticas. Para esto se indaga en el problema biológico y su interpretación en la bioinformática. A continuación los puntos del proceso que conforman el proyecto presentado son desarollados.

Por último se encuentran los resultados obtenidos por la detección de deleciones somáticas, se presenta además una evidencia que apunta a la presencia de las deleciones recurrentes entre distintos experimentos. De manera adicional, se realizó una comparación con genes relacionados con cáncer.

Capítulo 2

Marco teórico

En la actualidad las ciencias genómicas se han relacionado con una gran parte de las ramas de la biología, por esta razón es complicado definir de manera precisa a dicha disciplina. Con base en la raíz de la palabra "Genoma", puede entenderse que la "Genómica" es el estudio de los genomas. De cierta manera esta última definición involucraría a todas las ramas de la biología, dado que todos los aspectos de la vida se relacionan con su genoma.

La definición más concreta de la genómica, que es la más utilizada, es aquella que dice que se trata de una ciencia que estudia grandes cantidades de datos referentes a pares de bases en genomas y los métodos de alto rendimiento para su análisis (6). Por lo tanto, se puede afirmar que la genómica engloba la obtención y catalogación de ADN, así como la medición de su control transcripcional.

Posteriormente fueron adoptados nuevos paradigmas que explican la interacción molecular en los seres vivos, por lo que nuevas áreas de estudio surgieron de manera similar a la genómica. Así se generaron las ramas llamadas "ómicas", que estudian de manera parecida a la genómica la interacción de distintos tipos moleculares. Una de estas es la "Proteómica", la cual estudia la naturaleza y la interacción de las proteínas.

Otra rama de la biología que es necesario describir es la "Biología molecular", que estudia las consecuencias de las interacciones moleculares en los seres vivos. El dogma central de esta rama establece que el ADN es capaz de dar origen por un lado ADN, por medio de la replicación y por otro ARN, por medio de la transcripción, a su vez, el ribosoma puede realizar la traducción a partir del ARN mencionado a las secuencias de aminoácidos que componen a las proteínas. Este esquema explica la base del funcionamiento molecular de los seres vivos, así como la función de almacenar información del ADN.

Un genoma terminado es una secuencia de nucleótidos que representa la información contenida en el ADN de un organismo. Esto es útil para estudiar caraterísticas en los genotipos del mismo. Las tecnologías actuales de secuenciación son capaces de llevar a cabo lecturas en fragmentos de ADN con longitudes de hasta 1.8 Megabases (12). Por consiguiente, es imposible conocer el genoma de la mayoría de los organismos a partir de una sola lectura, por lo menos con las tecnologías actuales.

Para conocer la secuencia completa, los fragmentos, llamados lecturas, son ensamblados como un rompecabezas con piezas que empalman entre sí. Esta secuencia de lecturas empalmadas es llamada "contigs", los cuales van formando regiones de ADN que conforman a los genomas. Las regiones que no son posibles empalmar dentro de un cromosoma se llaman "gaps" (brechas).

Para que las secuencias consenso, incluidos contigs y gaps, puedan ser catalogadas como parte de un genoma terminado se requiere de una evaluación de calidad. Los altos grados de calidad permiten afirmar que las bases fueron identificadas correctamente. Para esto, es utilizada una puntuación llamada PHRED (9), que representa la calidad de descubrimiento para cada base detectada.

Una vez que se unen los contigs formados por lecturas de calidad, lo ideal sería que los gaps sean inexistentes. Sin embargo, esto no es así puesto que existen regiones de ADN que no pueden ser secuenciadas por métodos actuales debido a su naturaleza física o química. Además, existen límites en los contigs que no pueden ser resueltos sin ambigüedad. Una de las causas de este último problema son las regiones altamente repetitivas.

Una vez que se obtienen todos los contigs que se pueden resolver, se obtiene el genoma terminado, el cual puede servir de referencia para una especie o población. Cualquiera de las dos cadenas del cromosoma, en cualquier sentido puede representar el genoma. En ese sentido, es necesario seguir convenciones para evitar mantener representaciones uniformes de ubicación e información.

En cuanto a la información, las cuatro representaciones son equivalentes, pero no en cuanto a su interpretación biológica. La primera condición es un orden de 5' a 3', los cuales son carbonos que se encuentran en la desoxirobosa. Estos se usan de referencia, ya que son los que permiten conectar las bases nitrogenadas al conectar 5' con 3'. Este orden es el que se utiliza para codificar los aminoácidos que componen a las proteínas en las células. Esto se describe en el dogma central de la biología molecular. Como consecuencia quedan dos opciones, cualquiera de las dos cadenas, pero en sentido 5' a 3'. Usualmente se utilizará como referencia aquella que contenga el centrómero en las mínimas coordenadas, dejando una sola posible representación de referencia.

Los genomas terminados entonces pueden ser utilizados como referencia para identificar las regiones relevantes para una especie, ya sea por funcionalidad o por alguna característica presente. Los principales objetos de relevancia son los genes, estos son una secuencia de ADN que es capaz de ser transcrita, incluyendo los flancos del promotor, el cual inicia la transcripción y la región donde se detiene esta misma.

Toda esta información, que hace referencia a lugares específicos en el genoma, se llama "anotación", y es almacenada en bases de datos. Los genomas terminados de referencia se construyen por especie, lo que permite un consenso en la ubicación de las secuencias de nucleótidos para la misma (6). Tener un genoma terminado de una especie no es una imagen completa de la variación que esta puede presentar en distintos individuos dentro de su población. Por ello hay que tener en mente que la secuenciación de cada uno de estos puede contener variaciones que no están presentes en el genoma terminado que se usa de referencia.

Una referencia genómica para una especie permite mapear las lecturas de estas secuencias individuales a esta, así se obtiene el genoma terminado para el nuevo individuo, sin llevar a cabo el ensamble por medio de los contigs. El mapeo determina una ubicación probable de estas dentro del genoma de referencia. Idealmente el alineamiento entre ambas secuencias será perfecto, aún así, se reflejarán variaciones en las discrepancias entre ellas.

Tener una referencia de la estructura del ADN de un individuo no es tan representativo. Por lo tanto, fue necesario realizar sencuenciaciones en distintos individuos para obtener un consenso que resuelva regiones que no podrían ser ordenadas con un solo individuo. Sumado a ello, las limitaciones tecnológicas seguían produciendo gaps en el genoma. "El proyecto del genoma humano, en el 2001, resultó en más preguntas que respuestas" - (6). Pese a ello, conforme más conocimiento se ha adquirido, más información se ha aportado al genoma de referencia humano, así como de otras especies.

Actualmente, el genoma de referencia contiene gaps que no se han logrado resolver (6). La principal causa de esto es que hay regiones muy grandes con ADN repetitivo. Entonces los métodos actuales no nos permiten secuenciar y alinear contigs en dichas regiones sin ambigüedad. Esto se debe a que al alinear lecturas más cortas que dicha región, no exista un único alineamiento que sea viable.

Lo importante es recalcar que cada una de las lecturas pueden encontrarse en el genoma referencia, pero debido a la variación de cada individuo, presentará diferentes bases en la secuencia de nucleótidos. Dichas variaciones genómicas pertenecientes al organismo se llaman "Genotipos" (6). Las variaciones pueden ser desde un nucleótido, llamadas polimorfismos de un solo nucleótido (SNPs), o incluso de cientos o miles de pares de bases. En el caso del humano es de interés identificar dichos genotipos, ya que permite ubicar las mutaciones y relacionarlas con características presentes en el individuo.

Para tener una mejor imagen de toda esta variabilidad genética en individuos sanos, se creó el proyecto "1000 Genomes" (11). Dicho trabajo consiste en la recopilación de un gran número de genomas de personas provenientes de diversos grupos humanos en el planeta. Gracias a este esfuerzo se han logrado identificar variaciones poblacionales que no están relacionadas con enfermedades. De igual manera, el estudio provee una referencia para la identificación de enfermedades genómicas.

Debido a la gran extensión de los genomas, las tecnologías de secuenciación producen una gran cantidad de información, así es posible resolver preguntas del área biológica y genómica a partir de la información que se tiene de manera digital. Para lograr esto, es necesario desarrollar un software orientado a tales tareas.

2.1. Tecnologías de secuenciación

La secuenciación es extraer la información del genoma de alguna manera. Para poder hacer dicha tarea se realiza la secuenciación de distintas lecturas de forma paralela en un mismo dispositivo. De esta forma, la cantidad de recursos y tiempo es dividida

entre la cantidad de análisis en paralelo, que sean realizados en una secuenciación. Las tecnologías capaces de realizar dicha paralelización son conocidas como "NGS", por sus siglas en inglés Next Generation Sequencing (secuenciación de próxima generación).

Las máquinas que realizan la tarea de secuenciar ADN son producidas y desarrolladas por empresas privadas. Cada empresa cuenta con procedimientos distintos para la lectura de las bases. Al momento de hacer una investigación, los análisis tienen que adaptarse a las características de los datos que se proveen. La longitud de las lecturas, la baja calidad en los extremos y la inclusión de secuencias específicas en los extremos son unas de las posibles características que pueden presentarse.

Las tecnologías actuales son creadas y probadas por los mismos fabricantes, por lo que los usuarios no tienen control sobre las variables que influyen en la secuencia. A pesar de eso, los usuarios llevan a cabo procesos documentados para la producción de muestras, así como el análisis de los datos resultantes. Dichos procesos forman parte del conocimiento acumulado y documentado por la comunidad científica.

Esta homogeneidad incluye también el formato utilizado para los datos crudos obtenidos de las máquinas, independientemente de la empresa o laboratorio. Cualquier tipo de secuenciación produce un archivo en un formato conocido como "FASTQ", una variación de un formato llamado "FASTA". El formato FASTQ contiene secuencias de nucleótidos, junto con su respectivo valor de calidad llamado PHRED (9).

Las metodologías para obtener la información en lecturas varían dependiendo de la empresa, por ejemplo, son 454 Life Sciences, Illumina, ABI SOLiD, Ion Torrent, Pacific Biosciences y Oxford Nanopore Technologies. A continuación se describen estas para presentar los métodos.

454 Life Sciences

La tecnología de pirosecuenciación fue pionera en secuenciación masiva. El método consiste en fijar una sola cadena de los fragmentos de ADN a un extremo por medio de un cebador. Posteriormente se añaden bases nitrogeneadas de forma intercalada para sintetizar la cadena templada. Por cada nucleótido anclado se libera un pirofosfato. Como consecuencia se emite luz solo si se lleva a cabo la síntesis del nucleótido administrado. Este proceso se repite varias veces con el objetivo de poder tener un registro digital de los nucleótidos presentes en la lectura, el cual se almacena como un archivo en electrónico (5).

Illumina

La tecnología de esta empresa tiene un sistema similar a la piroluminiciencia. Allí bases nitrogenadas con marcadores fluorescentes de una longitud de onda caraterística son utilizadas para identificar a cada nucleótido. Lo característico del método es que realiza la lectura por medio de la síntesis de la cadena templada de ADN a la que se tiene fijada por un extremo. De esta manera se van añadiendo bases una por una, gracias a una terminación química en la base nitrogenada. Esta característica evita que se añada más de una base por ciclo de síntesis. Por lo anterior se requiere de una enzima que dé paso a la siguiente base en el proceso de síntesis. La lectura de luz se toma cada vez que una base es añadida a la cadena templada de ADN; este proceso consiste en una lectura controlada de las bases que se van añadiendo a la cadena de ADN (5).

Oxford Nanopore Technologies

Esta empresa desarrolló un dispositivo conocido como MinIon (27), que contiene un mecanismo molecular capaz de leer cambios en la carga a nivel de una base nitrogenada. El mecanismo hace pasar una sola hebra de ADN por un poro que aísla la polaridad de la base con respecto al ambiente. Por medio de esto es posible realizar una medición de cada nucleótido.

El proceso experimental consta de tomar un segmento de ADN de doble cadena. Uno de los extremos se polimeriza de manera que crea una sola cadena en conformación de horquilla. En el otro extremo se coloca un adaptador que le permite al poro detectar el inicio de la lectura. El proceso realiza una lectura de una cadena de nucleótidos y de su complemento, por un adaptador en forma de horquilla.

2.1.1. Bases de datos

El auge en las tecnologías de secuenciación generó una gran cantidad de información. Gracias al esfuerzo conjunto de centros de investigación se puede recurrir a ella desde cualquier parte del mundo. Este acceso a la información permite corroborar resultados, así como llevar a cabo investigaciones en una gran cantidad de muestras y especímenes.

El centro nacional para la información sobre biotecnología de los Estados Unidos "NCBI" junto con el instituto europeo de bioinformática "EMBL/EBI", así como el banco de datos de ADN de Japón "DDBJ", forman la colaboración internacional sobre la base de datos de secuencias de nucleótidos "INSDC"; todas estas siglas provienen de sus acrónimos en inglés.

Dichas instituciones proveen el almacenamiento y el manejo de la información de los estudios de secuenciación en el mundo. Así pues, contienen la información pertinente de cada uno de los experimentos que almacenan. En la actualidad la información entre las distintas instituciones está guiada por los lineamientos de la "INSDC", lo que permite su búsqueda y descarga, incluso entre institutos.

El "NCBI" tiene datos genómicos crudos almacenados en el archivo de lecturas secuenciadas, "SRA" por sus siglas en inglés. En dicha plataforma se encuentra la información de cada una de las secuenciaciones en el registro. En esta es posible descargar los datos crudos o alineados por medio de un formato comprimido llamado "cSRA". Dicho formato contiene la información necesaria en aras de poder extraer los archivos de secuenciación o alineamiento. Se debe mencionar que para el desarrollo de esta tesis se utilizaron experimentos provenientes de la base de datos mencionada.

Por su parte, "NCBI" ha desarrollado herramientas que permiten navegar en las bases de datos, descargar los archivos "cSRA", así como la extracción de información del mismo. Este software es llamado `SRA-Toolkit`, cuya documentación está disponible en la misma plataforma de internet. Con él es posible obtener los datos crudos de secuenciación, alineamientos, estadísticas y fenotipos, a partir de la información almacenada.

En primer lugar es necesario identificar el nombre del registro de interés en la base de datos. Con ese nombre se puede descargar el archivo con extensión "SRA", el cual

contiene toda la información necesaria. La descarga se puede hacer a través del servidor `ftp`. De igual manera, las herramientas informáticas que presenta `SRA-Toolkit` pueden extraer la información de dicho archivo de manera local a partir del "SRA".

Además de esta última, existen diversas bases de datos de distintas instituciones que permiten almacenar datos relacionados con los diversos campos de genómicas actuales. Los datos pueden ser lecturas crudas, alineamientos que fueron usados en los experimentos o anotaciones sobre los genomas conocidos. Debido a la diversidad de datos, distintos formatos se han desarrollado en pos de compilar la información relacionada con cada uno.

Con cada proyecto se generan nuevos datos que son reportados en los formatos pertinentes. Por ejemplo, el proyecto "1000 Genomes" está enfocado en identificar todas las variantes genéticas que presentan las distintas poblaciones humanas en el mundo (11). Por ese motivo, en su base de datos se puede encontrar desde la secuenciación hasta los genotipos identificados. También se pueden mencionar el Consorcio internacional de genomas de cáncer "ICGC" y el Atlas de genomas de cáncer "TCGA", los cuales están enfocados en unificar la información que se tiene en cuanto a las características genómicas del cáncer y las réplicas experimentales (5).

2.1.2. Mapeo de genomas

Una vez que se obtienen datos crudos de secuenciación, estos se componen de lecturas que deben ser localizadas dentro del genoma humano. Una lectura idealmente corresponde a una región definida del genoma de referencia, sin embargo, existe una gran cantidad de variaciones que pueden causar problemas. La secuencia de la lectura puede diferir en nucleótidos extra o faltantes, o en su defecto, el orden de estos. Para ubicar estas diferencias siempre se usa como referencia el genoma terminado de la especie. Por ello, buscar la región de la cual proviene la lectura consiste en buscar la región en la referencia en la que hay una mayor similitud con la lectura de búsqueda.

Para definir la similitud entre secuencias se han propuesto diversos modelos que penalizan distintas variaciones de la lectura con respecto a la referencia. Cada modelo permite la implementación de distintos algoritmos en aras de revisar la similitud entre secuencias; estos revisan cuáles partes de ambas secuencias se alinean entre sí. La similitud se traduce a una puntuación que se puede medir y comparar.

En esa medida, el problema que resuelve el mapeo es identificar la región de mayor similitud, es decir, la subsecuencia dentro de la referencia, cuyo alineamiento tiene la mejor puntuación. Dicho problema se conoce como el alineamiento local de secuencias, y puede ser resuelto por distintos algoritmos que varían en cuanto a eficiencia de tiempo y uso de memoria.

Un problema adicional son las lecturas con mapeo múltiple. Este caso se da cuando en una misma referencia con secuencias iguales o semejantes entre sí, una lectura puede tener la misma calificación en regiones distintas. Las regiones homólogas, mismas que provienen de un mismo gen ancestral, así como las repetitivas, son las regiones habituales en las que se presenta el problema en mención. Cabe anotar que en cada análisis

se decide y reporta cómo lidiar con dicho problema, con el fin tener observaciones comparables y resultados repetibles.

Asimismo, existe una clasificación de algoritmos llamados "de programación dinámica", los cuales también han sido utilizados para resolver el problema de mapeo. Estos manejan una representación matemática del problema y posteriormente dividen el problema en elementos más simples. De esta manera se utilizan cálculos sencillos para resolver el problema. Ejemplo de esto es el algoritmo de Smith-Waterman (57), que compara una secuencia de búsqueda con respecto a otra de referencia. Cada posible combinación de los caracteres entre las dos secuencias es representada por la casilla de una matriz.

Uno de los índices i de la casilla corresponde a la posición de un carácter de la referencia, y el otro índice j al de la secuencia de búsqueda. El valor inicial de cada casilla S es una calificación en función de la igualdad entre los caracteres de ambas cadenas, Q y R respectivamente, mostrado en la ecuación 2.1, donde a es la calificación que se atribuye a caracteres iguales.

$$S_{i,j} = \begin{cases} a & R_i = Q_j \\ 0 & R_i \neq Q_j \end{cases} \tag{2.1}$$

Posteriormente, cada casilla cambia de valor en base en las casillas aledañas. La ecuación 2.2 indica el nuevo valor $H_{i,j}$ que toma la casilla a partir de las casillas aledañas. De igual forma, la penalización del gap es representada con d. Las casillas que influyen están representadas en la ecuación 2.1.

Es así como se puede identificar el mejor alineamiento fundamentado en las casillas con el valor máximo de la matriz resultante. El mejor alineamiento es aquel que corresponde a la secuencia de casillas consecutivas con mejor calificación.

$$H_{i,j} = Max \begin{cases} 0 \\ H_{i-1,j-1} + S_{i,j} \\ H_{i,j-1} - d \\ H_{i-1,j} - d \end{cases} \tag{2.2}$$

Empero, se puede notar que en el caso de referencias grandes, el uso de memoria y procesamiento de datos es muy alto, debido a que se calculan todas las posibles comparaciones. Por consiguiente, se ha requerido desarrollar otros algoritmos que permitan identificar alineamientos con menos recursos.

Tal es el caso de la transformada Burrows-Wheeler (37), que permite un uso menor de procesamiento de datos y memoria, en la medida en que no requiere comparar todas las bases entre sí. El algoritmo consiste en construir un árbol de sufijos con todas las posibles subcadenas presentes en la referencia, que terminan en el final de la referencia. Esta estructura de datos permite identificar de manera consecutiva cada uno de los caracteres en una lectura. Las bases que se leen dentro del árbol, disminuye la cantidad de posibles alineamientos, por lo que se disminuye el número de comparaciones. De

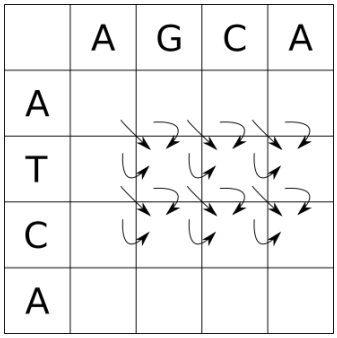


Figura 2.1: Representación del cálculo de las casillas en el algoritmo Smith-Waterman. A cada una de las casillas de la matriz $H_{i,j}$ se le tiene que calcular el valor por medio de la ecuación 2.2. En la imagen se ve la dependencia del cálculo de las casillas. La dirección de la dependencia permite un cálculo renglón a renglón en orden numérico de los índices, pero también muchas otras.

madera adicional, este enfoque permite paralelizar el trabajo de alineamiento, al realizar búsquedas de distintas lecturas al mismo tiempo en el mismo árbol de sufijos.

En consecuencia, el número de lecturas se puede repartir en hilos de procesamiento distintos. En este trabajo se utilizó `Segemehl` (24), que realiza un mapeo de lecturas exactas e inexactas. Hay distintos programas actuales capaces de llevar a cabo este proceso, pero este trabajo se centra con el uso de esta herramienta por consistencia.

`Segemehl` evita este problema al buscar los alineamientos locales de todas las regiones de una lectura. Para esto, un árbol de sufijos mejorado es utilizado, y representa todas las posibles secuencias ordenadas presentes en la secuencia de referencia (24). En ese sentido, es posible buscar cada una de las subsecuencias de las lecturas dentro de este árbol, y por ello, identifica alineamientos con inserciones y gaps.

Así pues, el funcionamiento de `Segemehl` consiste en buscar cada uno de los caracteres de una lectura en el árbol de sufijos y posteriormente las bases consecutivas. En virtud de que los segmentos de una lectura tienen coordenadas específicas, es posible identificar las inserciones y deleciones; mientras que las inserciones son bases no representadas en el árbol, las deleciones son saltos dentro del mismo. El proceso seguido por `Segemehl` es descrito en la Figura 2.2. Cabe mencionar que las lecturas mapeadas son el punto de partida para la mayoría de los análisis genómicos que se realizan en los individuos o especies, dado que representan el genoma total secuenciado. En este caso, se llevará a cabo un tipo de análisis de variación genética. A continuación se describirán los aspectos básicos de este proceso.

2.1.3. Análisis de variación genética

Un fenotipo es una característica, ya sea funcional o medible, que es el resultado del desarrollo de un organismo (22). Estos fenotipos son consecuencia de la interacción de los genes entre sí con los demás mecanismos de desarrollo e incluso del ambiente.

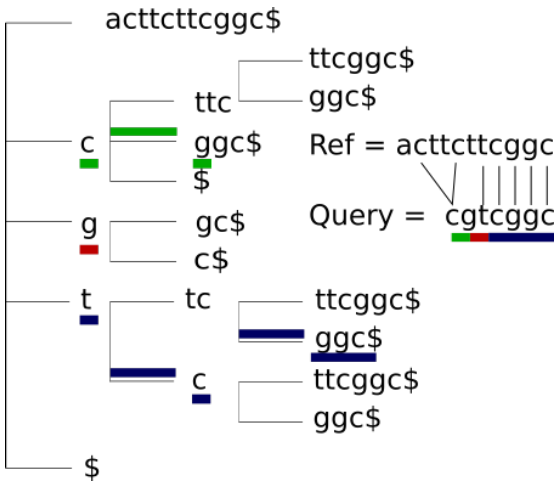


Figura 2.2: Árbol de sufijos implementado en `Segemehl`
En el árbol de sufijos se puede ver cómo la secuencia "acttcttcggc" se toma como referencia (Ref) para generar el árbol de sufijos que está desplegado. La secuencia "cgtcggc" (Query) es buscada en la referencia, la cual tiene distintas secciones alineadas en el árbol de sufijos.
La búsqueda inicia con la primera letra de Query, es decir la "c", la busca en la raíz del árbol, y busca si en dicha rama se encuentra el siguiente carácter, el cual es "g", peor esa rama ya no contiene el resto de la cadena, por lo que busca "g" otra vez en la raíz del árbol, ya que el siguiente carácter no es encuentra en las ramas que siguen, es decir "t" no sigue a "g", entonces se busca en la raíz nuevamente. Por último la subsecuencia "tcggc$" se logra alinear en las ramas subsecuentes y termina el algoritmo de búsqueda. El alineamiento más largo está en azul y permite encontrar una de las regiones con mayor similaridad, lo cual deja a los otros caracteres con dos posibles soluciones.
En el ejemplo "c" tiene dos posibles alineamientos o bien un mismatch; a su vez "g" solo puede alinear a las últimas letras "g", o bien hacer mismatch. Podemos ver cómo hay alineamientos alternativos en este ejemplo. Dependiendo de la configuración de `Segemehl`, se reportará alguna o las distintas alternativas. La imagen fue tomada a partir de un ejemplo en (24).

A partir de lo anterior suena tentador analizar la variación genética en busca de patrones que permitan identificar el rol de las mutaciones en distintos fenotipos. Esto resulta de interés en este caso en particular porque ayudaría a explicar o identificar las causas de distintos fenómenos. Unos de los fenotipos de mayor interés son las enfermedades genéticas, las cuales son consecuencia de mutaciones. Un ejemplo es la enfermedad de Sickle, la cual es producida por una variación en un solo nucleótido, a su vez causa una variación en la subunidad β de la Hemoglobina (6). Estas variaciones que pueden presentarse en un solo nucleótido de una población es llamada SNP por las siglas en inglés "Single Nucleotide Polymorphism".

Aunque hay relaciones directas como la antes mencionada, otros fenotipos pueden tener causas más complejas (6). Tal es el caso de la pigmentación de la piel, el cual es un fenotipo en el que influyen alrededor de 40 genes; entre más información se encuentra sobre el desarrollo de los organismos, así como de sus mecanismos moleculares, es posible

encontrar explicaciones a más fenotipos.

Las medidas llevadas a cabo por el consorcio de "1000 Genomes" indican alrededor de 84.7 millones de SNPs, 3.6 millones de inserciones o deleciones (indels) cortas, así como 60 mil variantes estructurales (11). Además, en el genoma de un individuo cualquiera pueden existir variaciones desde 4.1 millones a 5.0 millones de sitios con respecto al genoma de referencia (11).

La población con mayor variación genética, con alrededor de 5 millones de sitios por genoma, se encuentra en personas de ancestría africana (11), efecto que fue predicho con anterioridad por el modelo del origen africano de los humanos. Esto se debe a que las poblaciones africanas son más antiguas, entonces el tiempo permitió un mayor número de cambios en su genoma con respecto a los demás asentamientos humanos. Esto ilustra la gran cantidad de posibles cambios que se pueden encontrar al estudiar la variación genética a nivel de población.

Antes de llevar a cabo los estudios de variación, es oportuno tomar consideraciones relacionadas con aquello que se ha de analizar. Una de ellas es el enfoque para tratar regiones repetidas, en la medida en que estas pueden proveer ventajas o desventajas al momento de determinar variabilidad. Estos segmentos constituyen una fracción considerable del genoma; se conoce que en el humano, alrededor del 56 % está compuesto por algún tipo de ADN repetitivo (56, 58).

Como consecuencia de no considerar regiones repetidas, las variantes pequeñas pueden ser localizadas incorrectamente, a razón de que sus regiones aledañas pueden tener más de una región posible de mapeo (58). Por otro lado, regiones repetidas como las microsatelitales están relacionadas con una tasa alta de variación genética (6), cualidad por la cual las regiones repetidas pueden ser una herramienta de investigación para identificar individuos o grupos. Una vez que se obtiene la información del análisis, pueden ser inferidas relaciones entre el fenómeno estudiado y los genotipos. Formatos como GFF, VCF, BED, así como otros utilizados previamente en informática, son utilizados para reportar las relaciones encontradas.

En el estudio de variaciones celulares en un mismo individuo existen otras consideraciones. Por un lado, los métodos tanto experimentales como de secuenciación pueden inducir errores sistemáticos que pueden alterar las lecturas. Por el otro lado, el descubrimiento de variación es similar entre células clonadas y ADN replicado, e incluso tienen la misma fiabilidad.

La comprobación de similaridad fue estudiada por Brandon Milholland, et. al. (43), En este experimento se realizó la secuenciación genómica completa del ADN replicado a partir de células individuales. De manera simultánea, se llevó a cabo la secuenciación genómica completa de clonas de fibroblastos. Así se encontró que ambas secuenciaciones producían resultados similares, los cuales se comprobaron experimentalmente. Con la llegada de todas las tecnologías de secuenciación se produce una gran cantidad de información biológica nueva. Con ella se pueden contestar preguntas que antes estaban fuera de alcance. Asimismo, se da paso a nuevas incógnitas, por eso aún hay mucho que explorar en todas las ramas de la biología.

2.2. Biología del desarrollo

La biología del desarrollo es una rama de la biología que se dedica a examinar los fenómenos que suceden en los seres vivos durante el ciclo de vida. En el caso de los animales, este ciclo inicia con un óvulo fecundado, que contiene material genético proveniente de su ascendencia, suficiente para poder formar el organismo completo.

Sin embargo, cada uno de los organismos está compuesto por órganos a los que a su vez lo conforman tejidos que son construcciones de células con características diferentes. Estas últimas permiten implementar funciones específicas dentro del organismo, que son determinadas por la tarea que lleva a cabo el tejido y el órgano al cual pertenecen.

Una de las más grandes incógnitas que se ha intentado responder es cómo una célula puede dar paso a otras tan distintas. Esta pregunta se hace más intrigante con el hecho de que a excepción de unos pocos casos, las células de un organismo cumplen el principio de equivalencia genética, según el cual cada una de las células en un organismo tiene información genética suficiente como para conformar otro organismo igual. Es relevante notar que el principio no determina si el genoma es exactamente igual (61), por lo que sus detalles están siendo resueltos por medio de investigaciones de esta área de desarrollo.

El modelo actual de desarrollo indica que la especialización no es algo definido por la célula aislada, más bien es producto de una serie de cambios. Las células adquieren funciones y características distintas desde las primeras divisiones celulares después de la fecundación. En este punto de tiempo se forman tres capas principales en el embrión llamadas endodermo, ectodermo y mesodermo (61), a partir de estas se forman los demás órganos.

Pese a lo anterior, existen destinos específicos de especialización para cada capa. El ectodermo formará la piel y el sistema nervioso, el mesodermo se transformará en el esqueleto y el endodermo formará principalmente órganos internos. En los primeros estados de desarrollo, las células son capaces de cambiar en morfología y ubicación (61); estos cambios son coordinados gracias a la comunicación célula a célula. Por cada movimiento y división las células van especializando su función.

El proceso puede ser alterado por muchos factores, pero se mantiene controlado. Uno de los factores son modificaciones de ADN durante el proceso de desarrollo. Esto genera organismos mosaicos que contienen diferencias genómicas en sus células somáticas. Un caso similar se da con los organismos quiméricos (61); estos se dan como consecuencia de más de una célula fecundada. Cuando un organismo se puede formar con esta condición, entonces presenta grupos de células con genotipo distinto a las demás.

En ambos casos es posible desarrollar un organismo porque las células contienen la información suficiente para poder coordinar el desarrollo. Esto crea seres vivos con material genético distinto entre sus células somáticas. Es de destacar el hecho de que estas circunstancias también pueden traer consigo patologías en el organismo formado.

En esa línea, factores externos pueden modificar cualquier proceso de desarrollo (61). Por ello los mecanismos subyacentes tienen que ser lo suficientemente robustos para evitar que se forme un organismo aberrante a causa de una perturbación, en el

mayor de los casos. Esto se logra gracias a sistemas moleculares dentro de la célula, los cuales son redundantes y controlan la comunicación y expresión genética de la misma.

El desarrollo de los organismos vivos presenta muchas diferencias y se está en camino de aprender cómo funciona cada uno. Este proceso depende fuertemente de la información contenida en el ADN y de cómo esta misma se expresa. Entre más información se posea acerca de las modificaciones en el desarrollo del organismo, más conocimiento existe sobre los padecimientos que se pueden formar a consecuencia de los mismos.

2.2.1. Diferenciación celular

Este es el proceso responsable de los cambios en las células, desde un óvulo fecundado a un organismo pluricelular con órganos distintos. Este mecanismo es atribuido a los cambios de expresión de genes, lo cual es respuesta de distintos estímulos externos y estados internos en la célula. Dentro de la célula existe retroalimentación positiva y negativa, que consiste en promover o suprimir la transcripción de genes respectivamente. Por consiguiente, cambian las concentraciones de proteínas en la célula, ello causa cambios en la morfología o las sustancias secretadas por la célula. Es así como se va especializando a las comunidades de células de manera gradual en cada duplicación. Si bien la diferenciación celular puede estar acompañada de modificaciones en el genoma, su efecto o causa permanecen como una pregunta abierta. Por eso existe un área de investigación en torno a este tipo de mutaciones. Aun así, otros casos de mutaciones en las células somáticas son más estudiadas, como el caso de la poliploidía en el hígado de los mamíferos (19). La fusión de células hepáticas es parte del desarrollo normal de dichas células, por lo que adquieren genes duplicados, y por extensión, una mayor capacidad de expresión de genes.

2.2.2. Mutaciones Somáticas

Las variaciones genómicas en las células somáticas se han observado a distintas escalas, en varios momentos de desarrollo, tejidos y especies. Pueden presentarse como modificaciones en un nucleótido, hasta rearreglos o deleciones cromosómicas (15).

Las modificaciones pueden ser causadas por diversos factores, por ejemplo, la polimerasa es capaz de cometer errores al momento de la replicación, estos generan inserciones, deleciones o intercambios puntuales (15). Igualmente, hay influencia por parte de factores ambientales, mecanismos celulares, daño físico o químico, que son capaces de alterar las secuencias de ADN.

Es de destacar el hecho de que las mutaciones en general pueden causar tanto ventajas como desventajas adaptativas. Los genes están compuestos por exones, que contienen la información de las proteínas que transcriben, y por intrones, regiones intermedias entre los exones y no tienen información codificante. Por esta razón, las modificaciones en las secuencias dentro de los exones pueden generar proteínas aberrantes.

En contraste, mutaciones intrónicas pueden alterar de manera indirecta a la célula. Además de estas, existen otras regiones en el genoma que promueven o restringen la expresión de los genes. Estas alteraciones indirectas afectan la tasa con la cual los genes se expresan, y ello provoca cambios en la concentración de las proteínas.

De igual manera que la regulación genética cambia los fenotipos celulares, las modificaciones en el ADN puede modificar el funcionamiento de la célula. Por ello, las mutaciones somáticas se han relacionado con los padecimientos que se presentan durante el envejecimiento (60) o la carcinogénesis (32).

Una de las diferencias entre las mutaciones somáticas y las germinales es la cantidad de mutaciones que se presentan por división celular. En las líneas germinales hay tasas de 3.3×10^{-11} y 1.2×10^{-10} mutaciones por par de base por mitosis en humanos y ratón respectivamente. Empero, la tasa de mutaciones somáticas es de 2.66×10^{-9} y 8.1×10^{-9} en humano y ratón respectivamente. En este caso la diferencia es de por lo menos un orden de magnitud. Se han corroborado estos números por medio de estudios de secuenciación de ADN genómico completo, aplicado a células individuales amplificadas y a clonas de fibroblastos (43).

Los SNVs son variaciones que se descubren en comparación a otra referencia, cualquiera que sea. Por ejemplo, entre células de un mismo organismo. Gran parte de estas últimas se han reportado fuera de regiones exónicas (43). Esto puede ser atribuido a la extensión de las regiones intrónicas e intergénicas, o bien, a que las mutaciones en dicha región tiendan a ser silenciosas, por ende, no afectan la supervivencia de la célula.

Otra diferencia clave entre los dos tipos de mutaciones es la frecuencia que se presenta dentro de la población celular de un individuo. Las mutaciones germinales o poblacionales se encuentran presentes en todas las células, por el contrario, las mutaciones somáticas estarán presentes solo en una fracción de la población celular. Las somáticas pueden variar en proporción dentro de un tejido, lo cual depende del momento en el que ocurre el evento.

Solo podrán ser detectadas las deleciones somáticas si la fracción de células que contienen la mutación alcanza a ser detectada; esto depende de que la cantidad de células sea tan grande como para que exista una alta probabilidad de estar dentro del conjunto de células analizadas. Por consiguiente, el análisis de mutaciones somáticas se caracteriza por requerir un método, ya sea en laboratorio o in silico, el cual identifique las variaciones entre células de una misma muestra de tejido.

No se ha logrado comprobar alguna explicación ante la presencia de deleciones que ocurren en un tiempo y lugar específicos. Aquellas mutaciones somáticas, que son recurrentes entre individuos, dan indicio de que la mutación somática no es consecuencia del azar. Si existe algún patrón, entonces es probable que estén relacionadas con alguna característica conservada. Aun así, gran proporción de las mutaciones somáticas encontradas con las herramientas actuales, no son recurrentes entre individuos (49). En el experimento de OHuallachain (49) se encontraron 37 variaciones, de las cuales solo 2 fueron recurrentes en todos los individuos estudiados.

2.2.3. ADN circular extracromosomal

Las mutaciones somáticas pueden tener distintos efectos, por ejemplo, producir segmentos circulares extracromosomales de ADN "eccDNA" (por sus siglas en inglés Extrachromosomal circular DNA), a partir de regiones cortadas en el genoma (16). En su mayoría, estos eventos se encuentran en regiones repetitivas (45). Por lo anterior, han sido planteados mecanismos de corte basados en rearreglos genómicos por medio de secuencias homólogas.

Estas regiones de repetidos son capaces de integrarse al ADN cromosómico, por ello se ha visto que tienen el potencial de realizar traslocaciones (58). El eccDNA puede separarse del ADN genómico en aras de conformar segmentos independientes con capacidad transcripcional (16, 45). Posteriormente tienen la capacidad de integrarse al ADN, de nuevo por retrotranscripción. La función y el origen de dicho mecanismo es aún desconocida, no obstante, es posible que su presencia sea específica a tejidos.

Es de notar que esta relación entre deleciones y eccDNA fue reportada en tejidos sanos de cerebro, hígado y corazón en ratón durante su desarrollo, así como en líneas celulares de humano (17, 44). Ello en apoyo a la idea de la creación de eccDNA como consecuencia de la eliminación de segmentos en el ADN genómico. La longitud de dichos eventos no supera 1 kilobase y se presentan en baja frecuencia, además de que alrededor del 70 % de los eccDNA corresponde a regiones de repetidos.

2.2.4. Deleciones somáticas

Un subconjunto de mutaciones somáticas está compuesto por deleciones, que se han encontrado en lugares y momentos específicos del desarrollo en diversas especies y tejidos (10, 13, 14, 20, 21, 31, 34, 39, 46, 62). Sin embargo, su mecanismo y función siguen siendo preguntas abiertas. Podrían existir casos similares en los seres humanos. En caso de que estas deleciones estén presentes en el ser humano, tal cosa puede tener implicaciones sobre distintos padecimientos, tales como el cáncer (32) y aquellos que se presentan con el envejecimiento (6, 13), los cuales son procesos asociados con la acumulación de mutaciones.

2.2.4.1. Entre distintas especies

A continuación, se exploran animales en los cuales existe eliminación de una parte o de la totalidad del cromosoma como un evento programado. Encontraremos que pesar de no conocer bien la función de dichos eventos, es posible empezar a buscar eventos similares en el ser humano, incluyendo aquellos a menor escala.

La disminución de cromatina es un proceso programado donde cromosomas completos o secciones de ellos son eliminados de la célula. Anteriormente este proceso era considerado un fenómeno presente en pocas especies, pero luego se descubrió que diversos organismos de distintos filos han desarrollado y mantenido dicho proceso (34). Aun así, los detalles no han logrado ser aclarados del todo. Este proceso ocurre durante

el desarrollo normal del mismo organismo, incluso en lugares y momentos específicos. Generalizar este principio implica la presencia del mismo mecanismo en otras especies. No obstante, puede estar presente en una extensión menor del genoma.

Protozoarios ciliados

Estos son organismos unicelulares que tienen dos núcleos presentes; un núcleo se mantiene inactivo, y funciona como un núcleo germinal al transmitir la información genética a nuevas generaciones; el otro núcleo, por su parte, permanece activo y es el que funge de núcleo somático.

En un inicio, esto es, la fase latente de crecimiento del organismo, el núcleo somático es una copia del germinal, por eso contiene toda la información genética. Posteriormente es sometido a una serie de modificaciones que forman un núcleo maduro.

Este proceso inicia con la multiplicación de secuencias cromosomales repetidas que ya se encontraban presentes. En el siguiente paso ocurre una fragmentación de dichos cromosomas, llamados politénicos, debido a la anexión de dichas regiones. Los fragmentos, con tamaños similares a los genes presentes en el organismo, son encapsulados en vesículas. Por último, aproximadamente el 95 % de la cromatina es eliminada, lo cual deja aproximadamente el 5 % del material genético original, equivalente a 24 000 moléculas de ADN. EL núcleo pasa por una amplificación final, que genera 1 000 copias de cada fragmento restante (14).

Nematodos

El siguiente ejemplo se encuentra en los nematodos, principalmente `Parascaris` y `Ascaris` (46), los cuales se someten a modificaciones somáticas que son parte de su desarrollo. El primer tipo de modificación es la fragmentación de cromosomas. Esta es llevada a cabo en lugares específicos según el tipo de célula somática que es formada. Como consecuencia, se forman nuevos límites cromosómicos sin telómeros, y entonces inicia un proceso de anexión telomérica de zonas repetitivas de "TTAGGC" en los bordes fragmentados.

La evidencia sugiere que el mecanismo que sintetiza estas regiones no es el mismo que se encarga de la reparación habitual de telómeros. También puede ocurrir un proceso de eliminación de fragmentos o cromosomas enteros. Esto crea líneas somáticas con cromatina disminuida. En contraste, las líneas germinales mantienen la información genética y por último pasan por meiosis, tal como se muestra en la Figura 2.3.

Es necesario notar que las modificaciones deben ser prelocalizadas, de manera tal que se lleve a cabo un desarrollo completo. De lo contrario, las células podrían perder material genómico necesario para sobrevivir.

Este proceso de desarrollo se ha observado en 11 especies de nematodos, sin embargo, no es una característica conservada en todo el filo. Los nematodos que presentan eliminación de cromatina en fases presomáticas comparten un estilo de vida parásita. Un caso de peculiar interés en `Strongyloides Papillosus`, que tiene la facultad de desarrollarse como parásito o en forma libre pero solo presenta eliminación de cromatina en su forma parásita (46).

Se tiene la idea de que este mecanismo compensa la cantidad de genes duplicados que presenta el genoma de dichos nematodos (46). En este esquema la disminución de

cromatina permite regular una forma específica de expresión genómica para los tipos celulares, en vez de reprimir la expresión por otros mecanismos (20, 46). Esto contrasta con el hecho de que a gran parte de las secuencias de ADN que son desechadas la componen secuencias repetidas (34, 46).

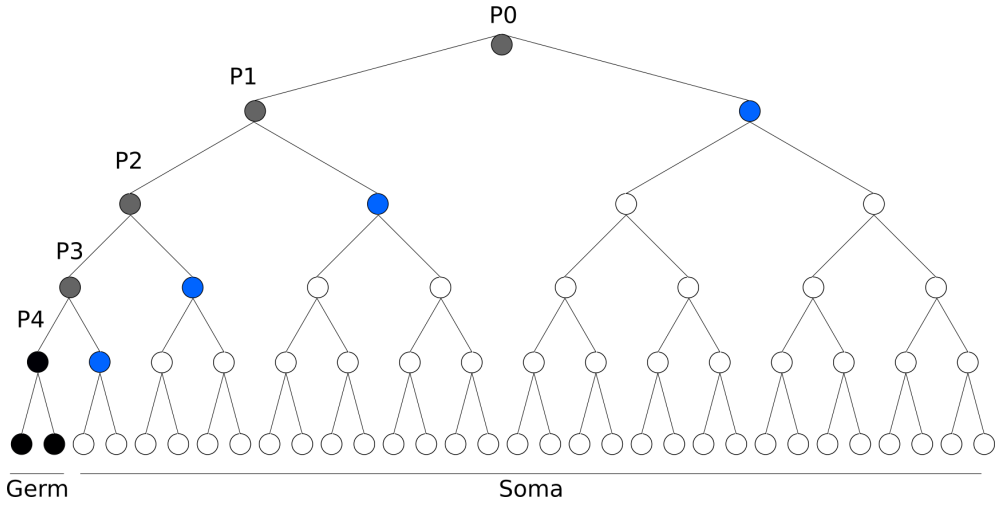


Figura 2.3: Destino somático de las células en el desarrollo de Nematodos
Esquema de proceso de eliminación de cromátida en nematodos. Células negras representan las células germinales que pasan por meiosis. Las células grises representan las células primordiales que contienen la información de zigoto. Las células azules son células que pasan por disminución. Las células blancas representan las células somáticas.

Crustáceos

Se encuentra igualmente una subclase de crustáceos diminutos que sufren mutaciones similares durante el desarrollo. En *Copepoda*, tres especies (*Cyclops divulsus*, *C. furcifer*, y *C. strenuus*) presentan pérdidas de aproximadamente el 50 % del material genético en células somáticas (14).

De manera similar, para *Mesocyclops* y *Cyclops kolensis* se encontraron valores de 90 % y 94 % respectivamente. En *Mesocyclops edax* se ha encontrado que gran parte del material genético que es eliminado en el proceso de desarrollo corresponde a regiones repetitivas satelitales, es decir, que en mayor parte corresponde a material genético que no transcribe (14). La disminución de cromatina ha aparecido de manera incidental a lo largo de distintos grupos evolutivos. Parece ser que RNAi interviene de manera similar que en los protozoarios ciliados (14).

Moscas

Estos eventos se pueden presentar en dos procesos del desarrollo; el primero es en las células germinales y el otro en las células presomáticas. El proceso completo es explicado con más detalle en la Figura 2.4. Los embriones provenientes de Diptera contienen más de un núcleo (21). Los núcleos germinales son aquellos que contienen toda la información genética del organismo. Los demás núcleos serán parte de las células

somáticas en los siguientes estadios.

La primera eliminación de material es el grupo de cromosomas designados como “E”, el grupo complementario designado como “S”. Existen variaciones del mecanismo, pero en general la eliminación toma lugar durante la mitosis. Los cromosomas designados a eliminarse se mantienen en un polo de la división. Por eso no son transmitidos a la descendencia celular (34).

El segundo evento consiste en la eliminación de cromosomas, que es el que de hecho define el sexo de la mosca. Esto se debe a que el sexo depende del número de cromosomas sexuales que tenga el genoma. A la par, las líneas germinales sufren una distinta eliminación de cromosomas, lo que da paso a gametos con distinta agrupación genética que las células germinales. Dependiendo del sexo, las células germinales producen óvulos a través de meiosis clásica, o bien, espermatozoides a través de un proceso de eliminación que descarta los alelos no sexuales pertenecientes al padre. A pesar de que todas las líneas celulares pasan por una reducción de cromatina, los gametos poseen una mayor cantidad de alelos en comparación con las células somáticas, en tanto presentan poliploidía.

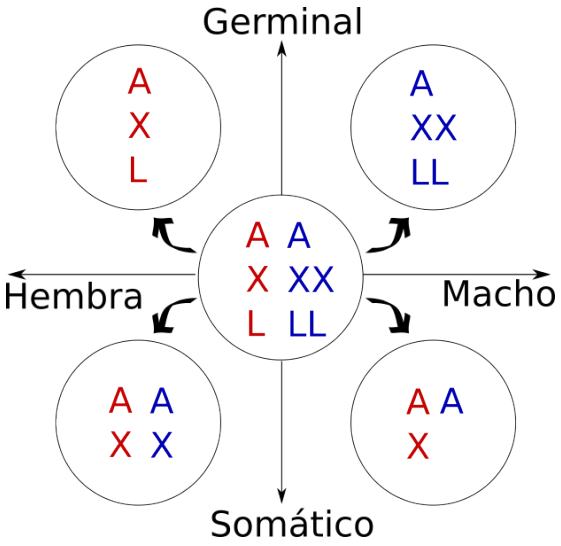


Figura 2.4: Modificaciones en los cromosomas en el desarrollo de Diptera
En el esquema se representa cómo es el conjunto de cromosomas que contienen las células somáticas y germinales en Diptera. Se observa la relación entre genotipo y sexo en el desarrollo. Además los cromosomas “L” se presentan solo en las células sexuales, pero son descartados en la definición del mismo. La célula del centro representa una célula fecundada con material genético del padre en azul, y de la madre en rojo. De esta manera podemos ver la forma en la cual los cromosomas se descartan de manera clara. Hay que notar que a pesar de que los espermatozoides tienen un solo color, el conjunto de cromosomas puede provenir del padre o de la madre pero no puede existir una combinación de cromosomas padre y madre en los espermatozoides.

Peces agnatos

Se tiene el caso de dos superclases de peces agnatos (sin mandíbulas), a saber, Cyclostomata y *Agnatha*, las cuales presentan eliminación de cromatina similar a la de los nematodos (34). Las células somáticas pierden de un 21 % a un 55 % de la cromatina total que se encuentra en las células germinales. En su mayoría es aquí eliminado ADN que contiene secuencias repetitivas, así como secuencias que son altamente conservadas entre distintas especies con el mismo mecanismo.

Marsupiales

En algunas especies de las familias *Peramelidae* y *Peroryctidae*, así como otros marsupiales presentan la característica de eliminar el cromosoma X o Y en distintos tipos celulares. Una posible correlación fue identificada entre las células que se dividen rápidamente y aquellas que pasan por deleción de cromatina (34).

2.2.4.2. Entre distintos tejidos

Ahora bien, las mutaciones también existen en tejidos específicos, pero su generalización a más especies permanece como una incógnita; pese a ello, aporta evidencia para buscar posibles formas de modificaciones a nivel somático.

Neuronas

Las células de las cuales se forman las neuronas, que se presentan en las primeras etapas embrionarias, pasan por un proceso de disminución de cromatina (13). Este proceso da paso a la neurola a partir de la blástula, presentando fases S de mitosis alargadas, debido a que se pierden regiones cromosómicas, rastro que será característico en el genoma cromosómico del sistema nervioso.

Las regiones microsatelitales y con replicones son las primeras en ser eliminadas. Dicho mecanismo se ha observado en *Cynops Pyrrhogaster* (62). Empero, grandes porciones de los cromosomas pueden perderse al grado de producir aneuploidia, como es el caso de los ratones (31). Como consecuencia se producen alteraciones en la expresión de genes.

Sistema inmune

Existen distintos tipos de compuestos ajenos a un organismo y tienen el potencial de causar algún daño. Cada uno de estos tiene componentes químicos que los identifican, por lo general son proteínas, también llamados antígenos, los cuales le permiten al mismo sistema reconocer los posibles patógenos de las demás sustancias.

El sistema inmune tiene la facultad de generar proteínas, a manera de molde de los agentes para que sean atacados o expulsados; así, tiene que ser capaz de sintetizar una gran cantidad de moldes para distintos antígenos (10). Para ello, las células del sistema inmune tienen la facultad de llevar a cabo rearreglos y recombinaciones genómicas de su ADN nuclear. Como consecuencia, son capaces de sintetizar una mayor diversidad de proteínas con una cantidad reducida de pares de bases.

Las células B utilizan estos rearreglos para generar proteínas llamadas anticuerpos, las cuales inician la eliminación del antígeno. Por el otro lado, las células T y B, especializadas para otros procesos de eliminación, utilizan esta plasticidad genómica

para sintetizar proteínas receptoras de membrana llamadas TCR's y BCR's respectivamente, con lo cual se obtiene la capacidad de detectar antígenos específicos. Todas estas proteínas tienen la tarea de anclarse a los antígenos debido a que su conformación funciona como un molde de estos últimos.

Unos cientos de genes Ig y TCR son capaces de sintetizar las proteínas necesarias. Estas modificaciones son de origen somático, de lo contrario, limitarían la capacidad de crear nuevos anticuerpos en la descendencia del organismo pluricelular. De manera específica, las mutaciones toman parte en las células B, las cuales son encargadas de mantener una memoria de los anticuerpos necesarios. Dichas células son capaces de vivir un tiempo más prolongado que las demás células del sistema inmune. Además la duplicación se lleva a cabo de manera que conservan células B con la capacidad de crear anticuerpos nuevos.

2.2.5. Envejecimiento

En cuanto al envejecimiento, este corresponde al proceso de desarrollo natural por el cual los organismos pasan desde su etapa adulta hasta la muerte. Distintos modelos permiten explicar dicho proceso, sin embargo, aún quedan muchas incógnitas. Entre estos el modelo de la *acumulación de mutaciones*, que propone que el envejecimiento es un proceso en el cual las células somáticas adquieren modificaciones en la expresión de distintos genes (6). Como consecuencia, los tejidos presentan los cambios característicos de edades avanzadas.

Se trata de un término poco entendido aún en estos días; se puede definir como el tiempo probable en el que un organismo se mantenga vivo desde un estadio de referencia (13), siempre y cuando el tiempo de vida sea interrumpido solo por causas aleatorias. Estas causas son la pérdida progresiva de las funciones fisiológicas, lo cual se conoce como *senescencia*. Las enfermedades que adquirimos pueden ser consecuencia de la aleatoriedad o de algún mecanismo subyacente, mismo que tiene como consecuencia el envejecimiento como total. Por el momento se supone que no existe un gen que codifique la muerte programada.

2.2.5.1. Los telómeros

Uno de los mecanismos candidatos para el proceso de senescencia es la pérdida de telómeros al ser un proceso vinculado directamente con la cantidad de divisiones celulares. Los telómeros son regiones que se ubican en los extremos del ADN y tienen la peculiaridad de estar constituidos por secuencias repetidas no codificantes (22). Al momento de replicar el genoma, las regiones de inicio de replicación no se encuentran exactamente en las últimas bases. Entonces, la síntesis de ADN no se realiza de manera completa, de esta forma acorta un fragmento reducido que pertenece al telómero.

Debido a que la región telomérica no codifica, el recorte no tiene efecto alguno en la expresión de genes. En caso de que la pérdida de información genética continúe de manera indefinida, entonces comenzará a dañar información que es relevante para

la célula. Así que es posible que la expresión de genes sea afectada, provocando una pérdida de las funciones celulares básicas y posiblemente la muerte de la célula.

La pérdida de telómeros es mitigada por un mecanismo de reparación de telómeros que minimiza la pérdida de material. Una enzima llamada "Telomerasa" contiene un templado de ARN de la secuencia telomérica, la cual puede anclarse a los extremos de los telómeros sin terminar. Posteriormente, la telomerasa indica la adición de una primasa en las tres últimas bases, lo que indica el final del cromosoma y evita que sea confundido como un segmento roto, y por lo tanto sea degradado. En un paso final, la polimerasa puede sintetizar la cadena de ADN restante.

La pérdida de la telomerasa repercute en efectos de senescencia en los organismos (22), debido a los cuales se ha propuesto que los telómeros y la pérdida de ADN en regiones contiguas puede tener una estrecha relación con el proceso de envejecimiento. Uno de los ejemplos más claros de esto es el síndrome de Werner (16), en el que la pérdida de los telómeros provoca un envejecimiento acelerado.

Igualmente, la pérdida de material genético puede tener efectos directos sobre la transcripción de proteínas, en caso de ocurrir en regiones codificantes. Aun así, el daño en otras regiones puede afectar de manera indirecta a la célula.

Todavía existe la posibilidad de que los telómeros cumplan la función de un reloj interno que provoca la senescencia. En definitiva, por el momento se puede afirmar que el envejecimiento es un proceso natural que hace parte del desarrollo de casi todos los organismos.

2.2.5.2. Teoría del desequilibrio genómico y epigenético.

Esta teoría es un modelo más general en cuanto a la pérdida de funciones a lo largo del tiempo (29, 39, 54); propone que ocurren cambios en las concentraciones de proteínas, que afectan el estado de las células. Los niveles de expresión de proteínas pueden ser causa del ambiente, enfermedades u otras circunstancias. Los cambios son efectuados sobre diversos mecanismos en la célula, los cuales pueden alterar las cadenas de ADN, ARN o las mismas proteínas. En el caso del material genético, el nivel de impacto depende principalmente de las regiones en las que sucede la modificación.

Otra forma de modificar la expresión es alterar los mecanismos epigenéticos, tales como la metilación en el ADN, modificaciones en histonas o cromatina, heredadas por duplicación dentro de las poblaciones de células somáticas, que entre otras cosas pueden cambiar a lo largo del tiempo.

Por otra parte, la expresión de los genes es regulada al interactuar dentro de una red con los demás genes y el ambiente. Como consecuencia, la síntesis de proteínas mantiene dosis específicas de estas. Pese a ello, alteraciones en las proporciones de expresión celular afectan dicha red de interacción, así como la forma en las funciones celulares.

En caso de existir mutaciones en las células, estas se van heredando en las células producto de su división; esto no evita que se creen nuevas mutaciones por cada nueva división, por lo que son acumuladas. Este efecto es conocido como el "tranquete de

Muller" (60), el cual ocurre en una población de organismos con reproducción asexual. En este caso, las células somáticas van acumulando mutaciones de igual forma, por lo que aumenta la frecuencia de mutaciones en función del envejecimiento.

La capacidad de producir dosis específicas de alguna sustancia depende de la información almacenada en el genoma y el epigenoma. Entonces el desequilibrio genómico consiste en el aumento de la probabilidad de corromper dicha información en función del envejecimiento, alterando las dosis de proteínas. El mecanismo presentado puede explicar padecimientos que se presentan a lo largo del envejecimiento.

En esa medida, hace falta determinar la causa de los cambios en el genoma. Una posibilidad es que existan eventos en lugares y tiempos determinados. También puede ser que las alteraciones sean producto de la aleatoriedad y de errores de duplicación. Ambos escenarios no son excluyentes. Hay que tomar en cuenta que efectos tales como la presión selectiva pueden causar eventos recurrentes en ubicación a partir de una causa aleatoria.

2.2.6. Cáncer

El cáncer es una patología que emerge por las células propias de nuestro cuerpo que empiezan a proliferar y competir con sus vecinas. Como consecuencia, las vecinas no son capaces de llevar a cabo sus funciones correctamente (32). Su creación y desarrollo está relacionado directamente con modificaciones en el genoma de las células en el cuerpo.

Estas modificaciones pueden tener dos efectos, uno es la activación de genes que dotan a las células de características carcinógenas, los cuales se llaman oncogenes. El segundo efecto es la supresión de genes que son capaces de frenar el comportamiento descrito, llamados genes supresores de tumores. Bajo este modelo, el cáncer es consecuencia de la acumulación de mutaciones somáticas que afectan ambos tipos de genes (32, 59).

Se han notado diversas modificaciones por las que pasa el cáncer, las cuales pueden silenciar genes supresores de tumores o causar una sobreexpresión de oncogenes. Unas de ellas son rearreglos genéticos complejos (cromoplexia, presente en próstata principalmente), deleciones, retrotranscripciones y amplificación por genomas circulares (en el caso de VPH) (59). En particular, la creación de eccDNA se caracteriza por un aumento en la expresión de oncogenes. Los cromosomas circulares contienen copias de oncogenes que pueden ser expresados de manera individual.

2.3. Software actual

Es importante mencionar que las variaciones genéticas pueden ser detectadas por una gran variedad de herramientas informáticas. Una de las aplicaciones más usadas es determinar genotipos en cáncer, para lo cual sirven programas como `SomatiCA` (8) y de `Control-FREEC` (3). También existen herramientas dedicadas exclusivamente a identificar mutaciones somáticas, tales como `POD` (2), que utiliza información de los

padres del individuo analizado. En general se requiere que la proporción de lecturas afectadas por la mutación sea grande, con fin de evitar artificios ocasionados por el ruido. A continuación se muestran otras herramientas que se relacionan con la búsqueda de deleciones en el ser humano que pueden ser de baja frecuencia.

2.3.1. GATK

El programa adquiere su nombre por las siglas en inglés "Genome Analysis Toolkit", que significa caja de herramientas para análisis de genomas. Este fue creado para poder identificar SNPs e indels germinales en secuenciaciones de exoma y genomas completos. También es capaz de llevar a cabo identificación de variantes de un solo nucleótido en células somáticas.

Para realizar la detección de variantes somáticas se requiere tener los datos de secuenciación mapeados. Posteriormente son detectadas las discrepancias que hay entre las lecturas y una referencia genómica. Al final se realiza un filtrado de los resultados, dicho filtro consiste en remover las regiones de las cuales no se pueden tener detecciones confiables de variación. Los filtros remueven regiones de baja calidad de alineamiento, regiones repetidas, variaciones germinales, contaminación, duplicados sistemáticos, regiones de baja cobertura, regiones definidas por el usuario, entre otras. El resultado es un reporte detallado de las variaciones que pasaron la prueba de los filtros. Así pues, `GATK` es una herramienta muy completa para la detección de variantes en experimentos donde se tiene acceso a una muestra de referencia.

2.3.2. SoloDel

Por otro lado, `SoloDel` (33) es capaz de detectar las deleciones somáticas con baja representación. El principal problema de dichas deleciones es su similitud con los efectos causados por el ruido. En vez de descartar las deleciones retirando regiones, los filtros de SoloDel se basan en el modelo de una mezcla gausiana que ambas deleciones, tanto somáticas como germinales, genera sobre las lecturas. En este punto el software requiere de otros programas de terceros, como `BreakDancer` (7) o `DELLY` (52), para realizar la detección inicial.

La entrada es un mapeo de lecturas pareadas. Estas se llaman así por su estrategia de seucuenciación, en él se llevan a cabo cortes de ADN en segmentos mayores a la longitud que es capaz de leer la secuenciación, pero lo más uniforme posible. En este caso el objetivo no es secuenciar todas las bases, pero solo las que corresponden a los extremos. Como resultado se obtienen parejas de lecturas que corresponen a regiones separadas por un espacio conocido, y cuyo orden de lectura es contrario entre sí.

Esta información permite detectar deleciones gracias a que la separación entre dos lecturas es conocida. Las deleciones son detectadas lecturas pareadas más lejos de lo que deberían. Posteriormente son eliminadas las lecturas que son parte de regiones de microrrepetidos, es decir, que tienen alta probabilidad de tener mapeos en las regiones aledañas. Después, por cada una de las localizaciones, en las que se encuentran posibles

deleciones, se calcula la probabilidad de que la profundidad encontrada en el intervalo sea causa de una deleción somática o germinal. En caso de que el primer tipo de deleción sea más probable, entonces la misma es reportada.

2.3.3. Sprites

`Sprites` (63) es un programa que descubre deleciones somáticas, cuyo principal interés es resolver deleciones dentro de regiones de microrrepetidos y microinserciones. Como entrada requiere lecturas pareadas mapeadas a un genoma de referencia. El programa supone que los puntos de corte de las deleciones siguen una distribución normal, por lo que une las deleciones que sigan esta distribución. Para lograr integrar la información, realiza realineamientos en las regiones con deleciones, de manera que se encuentren alineamientos alternativos que permitan reunir más deleciones al mismo conjunto. Esta herramienta permite identificar con mayor confiabilidad las deleciones en regiones homólogas. Sin embargo, su eficiencia de descubrimiento baja con profundidades de secuenciación altas y su tiempo de ejecución es alto debido a los realineamientos.

2.3.4. LUMPY

Es un programa basado en la idea de integrar las evidencias de deleciones por parte de las separaciones en lecturas pareadas, así como en las deleciones detectadas dentro de los mapeos de cada lectura. Las entradas corresponden a archivos de lecturas mapeadas de donde se obtienen las primeras evidencias. `LUMPY` (36) junta deleciones que son catalogadas iguales, acumulando la cuenta de cada una de las ubicaciones.

Reporta aquellas deleciones que pueden ser catalogadas como somáticas. Es posible excluir regiones específicas como parte de la ejecución del programa. También es capaz de utilizar distintos experimentos separados de manera simultánea para poder conjuntar evidencia. La capacidad de `LUMPY` en detección de deleciones es comparable con la de `Sprites` en cuanto a fracciones de uno en 10 cromosomas (63). La salida de `LUMPY` es una tabla con los parámetros de cada una de las deleciones detectadas.

2.3.5. DELLY

`DELLY` (52) es una herramienta que permite el descubrimiento de variaciones estructurales en los datos de secuenciación. También adquiere un mapeo como entrada, del cual es capaz de detectar las diferencias entre las lecturas y el genoma de referencia.

Cada una de las variaciones es representada como un nodo de un grafo. Una arista se traza en caso de que las variaciones se respalden entre sí. Las variaciones que sean parte de un clique (un subgrafo inducido completamente conexo) representan una sola variación, la cual es reportada.

Este método es uno de los primeros en ser desarrollado para resolver el problema de encontrar variaciones genéticas. No obstante, sigue siendo una referencia para com-

parar la eficiencia de métodos más recientes (36, 63) y su confiabilidad es alta para profundidades bajas de secuenciación de 5x a 30x.

2.3.6. MosaicHunter

`MosaicHunter` (25) es una de las herramientas más actuales que permite llevar a cabo el descubrimiento de variaciones genéticas por medio de experimentos de secunciación tanto genómicos como exómicos. Este programa determina cuál es la probabilidad posterior de que esta variación esté presente en un grupo de células, a manera de una población mosaico de células. Incluso, el programa es capaz de determinar la probabilidad de que las variaciones tengan un origen germinal, El programa integra imformación de posibles errores de secuenciación, alineamiento y la información de variación alélica reportada en dbSNP.

Al final se presenta un listado de todos los nucleótidos que presentan algún tipo de variación, junto con sus características, así como los valores de probabilidad y fiabilidad de pertenecer a alguna de las categorías siguientes: germinal homozigota, germinal heterozigota o somática mosaico.

2.3.7. VarScan2

`VarScan2` (35) es un programa que busca alteraciones somáticas y variciones en el número de copias. El resultado de la búsqueda son variaciones de un solo nucleótido (SNV's) los cuales son evaluadas como somáticas en base a la profundidad local de las regiones con variación. Sin embargo, este proceso es utilizado para analizar células de poblacionales de cáncer, por lo que es necesario tener un experimento de referencia. Por el otro lado, `VarScan2` puede identificar SNPs e indels en experimentos individuales.

Capítulo 3

Análisis de deleciones somáticas en el genoma humano

Es de suponer que las deleciones somáticas están presentes en tejidos de un individuo formado. Si se da uno de estos eventos en una ubicación específica ocurre que una fracción de células en un tejido lo contendrá en su genoma. Para tener un indicio claro de estas variaciones se puede analizar cualquier experimento de secuenciación realizado con anterioridad.

El descubrimiento de deleciones se puede hacer por medio de (2, 3, 8, 33, 36, 52, 63). Estas herramientas descartadan observaciones que pueden asemejarse a errores sistemáticos. El objetivo es obtener todas las deleciones presentes en distintos individuos, con el fin de compararlas con otros experimentos y verificar si ocurren en una ubicación específica.

Las regiones repetitivas son fuente de problemas, entre ellos se pueden nombrar la ambigüedad de alineamiento y el diseño de cebadores con múltiples anclajes. Bajo esta idea, fueron excluidas las deleciones que se localizan dentro de secuencias repetidas en el genoma de referencia.

Una descripción gráfica del objetivo de este trabajo se presenta en la Figura 3.1. Los detalles técnicos de las herramientas implementadas en esta tesis están descritos en el apéndice. De igual manera, el repositorio cuenta con la documentación necesaria para su uso o adaptación. A continuación son descritos los pasos llevados a cabo para la obtención de las deleciones de origen somático.

3.1. Obtención de datos

Para evitar que los métodos de curado y mapeo introduzcan más variables al análisis, se realizó el análisis a partir de los datos crudos. El primer paso fue identificar los experimentos de los cuales se pueden obtener datos crudos de secuenciación para analizar.

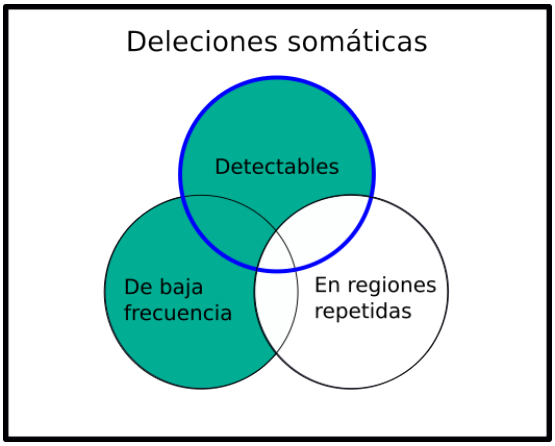


Figura 3.1: Diagrama de Venn que explica las deleciones de interés.
Consideremos el universo de deleciones somáticas, en el cual hay tres conjuntos que se sobrelapan. El primer conjunto es catalogado como "Detectables", porque se pueden obtener por otros programas. El segundo grupo son las de "Baja frecuencia", las cuales son difíciles de detectar; el último grupo corresponde a aquellas que se encuentran en regiones de repetidos. Mientras que las herramientas actuales permiten detectar deleciones que se encuentran en el círculo azul, nuestro interés son aquellas que se encuentran dentro de la región sombreada.

3.1.1. Búsqueda de datos de secuenciación

Pese a que en las bases de datos existe una gran cantidad de experimentos, estos pueden contener características que no son de utilidad para el objetivo de este estudio. De tal modo que la primera tarea fue determinar el conjunto de experimentos a utilizar por medio de los criterios descritos a continuación.

Para lograr identificar variaciones en cualquier región del genoma, este tiene que estar completamente contenido en las lecturas. Por ello se utilizan datos provenientes de la estrategia de secuenciación genómica completa "WGS". También fue necesario corroborar que en efecto esta estrategia contuviera al genoma nuclear completo.

Fue necesario considerar que las bases de datos tienen reglamentos en cuanto a la disponibilidad de datos genómicos humanos. Fue por eso que se enfocó en aquellos proyectos de acceso público, con el fin de no requerir permisos adicionales para descargar datos.

Debido a que se buscaban eventos que fueran parte del desarrollo normal del ser humano, las células analizadas debían estar libres de modificaciones, las cuales sean consecuencia de padecimientos o de la metodología experimental. En esa medida, la búsqueda fue enfocada en proyectos que analizaran tejidos sanos y que utilizaran procedimientos que no afectaran el ADN.

Como último criterio se tomó en cuenta la profundidad de la secuenciación. Este parámetro se refiere a la cantidad de lecturas que cubren cualquier coordenada en el genoma en el caso ideal de que estas tengan mapeos distribuidos uniformemente. El

cálculo para obtener esta medida es dividir la cantidad de bases secuenciadas entre la longitud del genoma.

La profundidad es un aproximado, puesto que en la realidad las lecturas no se distribuyen uniformemente. Esta medida permite aproximarse a la resolución de detección de deleciones; a mayor profundidad, mayor población celular representada, y por lo tanto, un evento de baja frecuencia tiene mayor probabilidad de ser representado.

Las bases de datos exploradas en un inicio fueron "ICGC" y "1000 Genomes project"; sin embargo, "ICGC" no contiene experimentos de secuenciación de acceso público. En cuanto a "1000 Genomes project", contiene experimentos de 30X, considerada alta profundidad, por lo que podría ser útil para este tipo de análisis. Sin embargo, las secuenciaciones en este proyecto son sobre líneas celulares son inducidas pluripotenciales. Estas líneas se obtienen al modificar genómicamente el ADN de las céluas, por lo que estos datos no fueron utilizados para este proyecto.

La base de datos más completa y con posibilidad de encontrar secuenciaciones realizadas sobre células sin modificar fue SRA. Por lo tanto, el primer paso fue buscar experimentos de secuenciación masiva dentro del acervo de SRA. En la búsqueda se aplicaron las características indicadas anteriormente. Más adelante fueron detectadas aquellas secuenciaciones que provienen de tejido sano sin modificaciones genómicas, en tanto que esta restricción no se logró discriminar automáticamente.

En un inicio se identificaron grupos de secuenciación de sangre y de seno, cuatro de cada uno, así como dos más de próstata. Sin embargo, al final fueron elegidos cuatro experimentos secuenciación con mejor calidad de lecturas, tal y como se presentan en la Tabla 3.1. Los experimentos de próstata eran muy grandes en volumen de datos, por lo que procesos como la descarga y mapeo posteriores presentaban errores de procesamiento, así pues, fueron descartados para este análisis.

Nombre	***Sexo***	***Procedencia***	***Bases (Gb)***	***Número de corridas***	***Grupo***	***Año***
SRX257065	F	Seno	426	14	Stanford University	2013
SRX257088	F	Seno	213.7	10	Stanford University	2013
SRX1660320	M	Sangre	249.4	2	Chulalongkorn University	2016
SRX237626	M	Linfocitos	86.1	3	The Genome Center at Washington University	2013

Tabla 3.1: Tabla de experimentos utilizados en el análisis.
Cada uno de los cuatro experimentos, utilizados para esta tesis contiene su precedencia biológica y las bases secuenciadas en total, las cuales se reparten en el número de corridas.

Cada experimento corresponde a una muestra muestra biológica, la cual fue secuen-

ciada un determinado número de veces. Cada una de estas secuenciaciones de la misma muestra es una corrida en la Tabla 3.1. En un experimento, su cantidad de bases es la adición de las que componen cada una de las corridas.

3.1.2. Descarga

Una vez identificados los experimentos en SRA, estos fueron descargados por medio del servidor FTP de NCBI. Las lecturas secuenciadas en formato FASTQ fueron almacenadas en el servidor de ADN en el LAVIS (UNAM).

3.1.2.1. Formato FASTA y FASTQ

En el formato FASTA cada lectura tiene un identificador, dicha cadena de caracteres es identificada por el carácter “>” al inicio. La cadena en la siguiente línea corresponde a la secuencia de nucleótidos. La implementación del formato fue desarrollada por Bill Pearson (50). Posteriormente, la información de calidad fue añadida, de manera que se formó el actual FASTQ (9), en el que se introdujeron dos nuevas líneas. La primera línea es el identificador con el carácter “+” al inicio, lo que indica que la cadena de la siguiente línea la componen caracteres ASCII, que representan los valores de calidad. En consecuencia, las cadenas de calidad y nucleótidos tienen la misma longitud. Además, el símbolo “>” de la secuencia de nucleótidos se reemplazó por el carácter “@”.

3.1.3. Calidad de experimentos

Las máquinas de secuenciación pueden presentar baja calidad en las lecturas debido a la incertidumbre de los sensores. Así que es necesario revisar posibles errores, y después corregirlos en pos de evitar desarrollar el trabajo con información imprecisa. Cabe destacar que fue posible revisar la calidad de las lecturas por medio de `fastqc` (5). De manera posterior, fueron descartadas las bases que no pueden aportar información confiable al alineamiento. La edición fue realizada por medio de una herramienta llamada `fastq_quality_trimmer` de “fastqx-Toolkit”. Empero, aún existían calificaciones que no permitieron utilizar algunas corridas, lo cual fue también revisado por medio de `fastqc`. Es por ello que se descartaron las corridas que no pasaron las pruebas de calidad después de la edición.

3.1.4. Mapeo de lecturas

Luego del curado de datos crudos se procedió al mapeo. Para esta tarea fue utilizado el programa `Segemehl` (24). La configuración aplicada permitió alineamientos con lecturas separadas (“Split reads” en inglés), para poder ubicar las deleciones de mayor longitud.

3.1.4.1. Formato SAM

El formato SAM contiene representadas las bases y su alineamiento con respecto al genoma de referencia, llamado así por las siglas en inglés "Sequence Alignment/-Map", que significa "formato de alineamiento o mapeo de secuencias". Cada campo del formato está delimitado por tabuladores entre los campos de cada renglón. El grupo compuesto por las primeras líneas es llamado "encabezado"; esta es opcional y tiene la característica de iniciar las líneas con un carácter de "@". El encabezado indica el tipo de ordenamiento o agrupamiento que contiene el archivo, la información del archivo que se usó como referencia para el alineamiento, así como de cada uno de los grupos que componen al alineamiento.

En la siguiente sección del formato se encuentran las secuencias mapeadas. Cada uno de los registros contiene 11 campos que permiten conocer la forma en la que la lectura se alinea a una referencia. La descripción de cada campo se encuentra en la Tabla 3.2.

Número	***Nombre***	***Descripción***
1	QNAME	Nombre de la lectura que corresponde al indicador en el archivo de datos crudos.
2	FLAG	Un valor cuya representación binaria indica qué propiedades tiene la lectura mapeada.
3	RNAME	Nombre de la secuencia de referencia.
4	POS	Posición del CIGAR que comienza a corresponder a bases del genoma de referencia.
5	MAPQ	Valor logarítmico de la calidad del mapeo. Un valor de 255 indica que la lectura no se puede utilizar.
6	CIGAR	La codificación de la forma en la que las bases de la lectura se alinean en la referencia.
7	RNEXT	En el caso de las lecturas pareadas, este campo indica el nombre de la lectura siguiente.
8	PNEXT	En el caso de las lecturas pareadas, indica la posición de la siguiente lectura.
9	TLEN	En el caso de las lecturas pareadas, indica la distancia entre los extremos opuestos de la región generada por ambas lecturas.
10	SEQ	La secuencia de nucleótidos de la lectura.
11	QUAL	La calidad de las bases de la lectura, tal como se indica en los FASTQ.

Tabla 3.2: Tabla de campos en el formato de lecturas del archivo SAM
Descripción de cada uno de los campos que tiene cada registro del formato SAM. La primera columna indica el número de campo, la segunda el nombre abreviado que se utiliza en la documentación y la tercera es la descripción.

CIGAR

Este campo en específico es el que se utilizó para detectar las deleciones. En él es indicada la forma en que la lectura es alineada con respecto a la secuencia de referencia. La codificación consiste en una cadena de caracteres, que consta de números seguidos de una letra. El número indica la cantidad de bases consecutivas con la descripción de la letra, las cuales están expuestas en la Tabla 3.3.

Cada una de las descripciones puede corresponder a un carácter en la referencia, la lectura, ambas o ninguna. Solo se puede aplicar una de estas relaciones a cada base en cada secuencia, por lo que se dice que "consume" elementos de alguna cadena. Dicho consumo está representado en la Figura 3.2. En caso de no existir un alineamiento, el CIGAR es sustituido por el carácter "*".

Un ejemplo es presentado en la figura 3.3. En ella se observa como es que el CIGAR representa cada uno de los caracteres en una lectura. Los caracteres "N" se utilizan como deleciones, pero por lo general indican que una región falta debido a el método utilizado para secuenciar. Por ejemplo, secuenciando exomas, siempre faltarán regiones intrónicas. Por último, cabe mencionar que los caracteres "S" y "H" no se presentan en el ejemplo, estos se refieren a cadenas parte del anclaje utilizado para llevar a cabo el proceso de secuenciación, por lo que a veces se encuentran presentes dentro de la secuencia de la lectura, sin embargo para este trabajo fueron ignoradas.

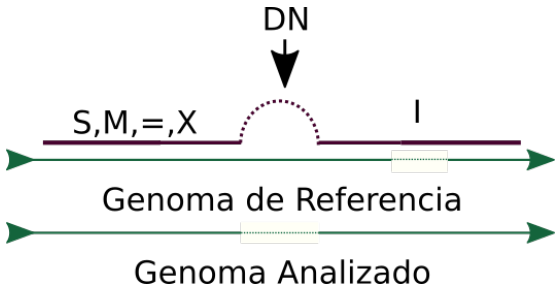


Figura 3.2: Descripción del consumo de secuencias CIGAR
El CIGAR es la representación de la forma en la cual la lectura se alinea con el genoma de referencia. Los caracteres "=,X,M" indican que existen nucleótidos consecutivos en la lectura que corresponden a la referencia. Los caracteres "D" y "N" indican el gap, que presenta la lectura. Finalmente el carácter "I" indica la presencia de una base en la lectura que no se encuentra en la referencia. Las líneas continuas representan regiones consumidas por una característica, de manera contraria que las punteadas.

3.2. Detección de deleciones

El alineamiento de las lecturas sobre el genoma de referencia provee la información necesaria para identificar deleciones. Antes de la detección se filtraron aquellas deleciones con una calidad de mapeo `QMAP` menor o igual a 5.

Carácter	*Descripción*	*Consumo de lectura*	*Consumo de referencia*
M	Indica el alineamiento de la base de referencia con la lectura, aunque no sean la misma base.	Sí	Sí
I	Indica la inserción de una base presente en la lectura, pero no en la referencia.	Sí	No
D	Indica una base eliminada en la lectura, por lo que sí aparece en la referencia.	No	Sí
N	Indica una región que se salta en la referencia.	No	Sí
S	Indica la región de recorte suave presente en la lectura.	Sí	No
H	Indica la región de recorte fuerte que no está presente en la lectura.	No	No
P	Relleno (región que no está presente en ninguna secuencia).	No	No
=	Indica el alineamiento de la base de referencia con la lectura, donde la lectura y la referencia contienen la misma base.	Sí	Sí
X	Inicia el alineamiento de la base de referencia con la lectura, donde la lectura y la referencia no contienen la misma base.	Sí	Sí

Tabla 3.3: Tabla de caracteres del CIGAR en el archivo SAM
El CIGAR tiene caracteres que representan distintos atributos del número de bases que lo preceden. En la tabla se expone el significado de cada carácter. El consumo de una secuencia indica si el atributo aplica a bases en la lectura o la referencia.

Es oportuno señalar que fue necesario agrupar los gaps que eran semejantes dentro de un mismo experimento. En el siguiente paso fueron identificadas las variaciones que tenían un origen somático y por último, se determinó el subconjunto de variaciones recurrentes entre los individuos.

Dicho lo anterior, a continuación son descritos los pasos que llevaron a la detección de deleciones somáticas. En algunos casos, las tareas requirieron la implementación de herramientas propias.

3.2.1. Lectura de CIGAR

A cada una de las lecturas les fueron buscadas las deleciones contenidas en ellas mediante el CIGAR. Además, fue necesario reducir lo mínimo la probabilidad de que la deleción sea un artificio debido a las ambigüedades de mapeo. Una forma de lidiar con este problema fue crear la condición de anclaje, que consiste en una cantidad mínima de bases presentes en la lectura a ambos lados de una deleción. En la Figura 3.4 se

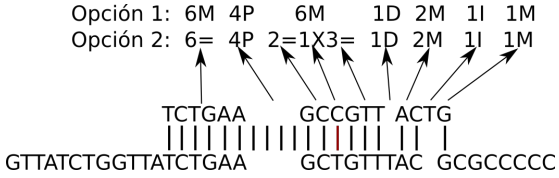


Figura 3.3: Descripción del CIGAR
En este caso tenemos dos secuencias, una inferior y otra superior que corresponden a la referencia y a lectura respectivamente. Las flechas indican de donde viene cada parte de la representación, la opción 2 es la que está mas detallada, ya que indica los tipos de alineamientos por medio de los caracteres "=" y "X", pero pueden ser reemplazados por una descripción más general con el carácter "M". Esta otra posible representación se observa en la opción 1.

representa cómo son discriminadas las deleciones.

Es posible clasificar las regiones del CIGAR en dos grupos; el primer grupo lo conforman las bases que aportan al gap, que corresponden a los caracteres "N" y "D"; el segundo grupo lo constituyen bases que aportan al anclaje, representadas por las letras "I,=,X,M", y "S".

De manera específica, la tarea es determinar las longitudes de los gaps y anclajes de cada una de las deleciones presentes en los mapeos. Para lograr esta detección, fue implementado un programa llamado `SAM_parce_GAP`. La cantidad mínima de bases para identificar una región sin ambigüedad es por lo general de 20 bases (53), por lo que se utilizó este valor como mínimo de longitud en los anclajes.

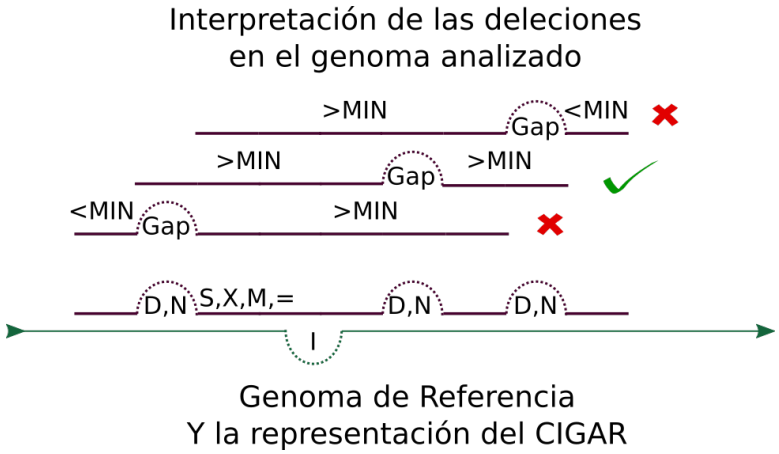


Figura 3.4: Descripción del algoritmo de interpretación de CIGAR
La lectura es representada como la última línea en morado se alinea a la referencia, de la cual es obtenida la información de tres posibles deleciones reportadas, las cuales son representadas con las líneas de morado superiores. Sin embargo, un mínimo de longitudes en los anclajes es necesario para reportar una deleción, por lo que solo la segunda línea es reportada.

Algoritmo

El algoritmo recorre cada una de las lecturas alineadas por medio del CIGAR. El alineamiento de una lectura es analizado en pro de obtener la cantidad total de bases que aportan como anclaje, este valor es almacenado como el segundo flanco. Posteriormente se analiza el CIGAR a través de un ciclo que detecta los gaps a fin de evaluar las longitudes mínimas de gap y flancos en el ciclo de análisis que sigue.

Este ciclo primero almacena una región consecutiva que no contenga algún gap, al que se le llama *delta de flanco*, ello en vista de que al encontrar el gap es añadido al flanco izquierdo aunque sustraída al derecho. Después es almacenada la cantidad de bases consecutivas que forman un gap. Si los anclajes de la deleción superan la longitud mínima requerida entonces es reportada. Posteriormente se repite el ciclo, mismo que se detiene cuando se terminan los elementos por analizar en el CIGAR.

3.2.1.1. Formato BED

El formato BED, del acrónimo en inglés "Browser Extensible Data" que significa "Datos Extensibles para Navegadores", consiste en texto ASCII separado por tabuladores. Los campos más representativos son el cromosoma y dos coordenadas dentro del mismo. Dicho formato es útil para representar información de intervalos dentro de los genomas, los cuales están descritos en la Tabla 3.4.

Número	***Nombre***	***Descripción***
1	Cromosoma	Nombre del cromosoma donde se encuentran las coordenadas.
2	Coordenada de inicio	Inicio del intervalo. La primera coordenada en el cromosoma siempre es 0.
3	Coordenada de fin	Fin del intervalo. Indica que la última coordenada dentro del intervalo igual a su valor menos uno.
4	Nombre	Nombre que identifica dicha línea dentro del BED.
5	Grado	Una calificación de 0 a 1000, la cual indica que tanto se califica un rasgo en la región.
6	Cadena	Identifica el sentido de la cadena de ADN con respecto a la referencia. Puede ser "+" o "-".
7	Inicio Grueso	Indica la coordenada de inicio de una característica dentro el intervalo.
8	Fin Grueso	Indica la coordenada de fin de una característica dentro el intervalo.
9	RGB	Un campo de tres números separados por comas que corresponden al código RGB de un color. Dicho color será aplicado al intervalo correspondiente a la línea del BED.
10	Cuenta de bloque	Es el número de bloques descritos en los siguientes campos.
11	Tamaño de bloque	Es texto separado por comas que indican la longitud de cada uno de los bloques dentro del intervalo.
12	Inicio de bloque	Es el texto separado por comas que indican las coordenadas de inicio de cada uno de los bloques. Este campo tiene que corresponder en cantidad de elementos con el anterior.

Tabla 3.4: Tabla de campos en el formato de lecturas del archivo BED
Descripción de cada uno de los campos que tiene cada registro del formato BED. La primera columna indica el número de campo, la segunda el nombre del campo, y por último, la explicación.

3.2.2. Filtrado de datos

En virtud de que la presencia individual de deleciones no contiene la suficiente información para identificar orígenes somáticos, es necesario conjuntar las deleciones semejantes. Antes de esto se requiere filtrar las deleciones que están cercanas a regiones de repetidos. En última instancia, se debe descartar a las que son de origen germinal con la información conjuntada.

El primer filtro consiste en remover lecturas cercanas a regiones de repetidos en el genoma. De esta manera se evitaron deleciones provenientes de artificios en el mapeo, así como problemas en el diseño de cebadores. Después fueron identificados los grupos de deleciones que se sobrelapaban, con lo que fueron agrupadas.

Para separar las deleciones de origen germinal fue calculada la fracción de lecturas que contenían el mismo gap en un experimento. Se clasificaron como somáticas las deleciones que no estaban presentes en el 50 % o 100 % de las lecturas. Esto fue llevado a cabo de esta manera porque consideramos la posibilidad de eventos de corte que sean tan frecuentes que afecten ambos cromosomas en una célula, pero no en todas. A continuación son descritos cada uno de los procesos, en el orden en que fueron aplicados.

3.2.2.1. Regiones repetidas en RepeatMasker

Los alineamientos en regiones de repetidos en el genoma carecen de fiabilidad, dado que un flanco puede presentar ambigüedad de mapeo en alguna región consecutiva. De forma adicional, un cebador diseñado dentro de regiones de repetidos puede anclarse a distintas coordenadas y ello representa retos adicionales para la comprobación experimental de deleciones en dichas ubicaciones. Una gran cantidad del genoma consiste en regiones de repetidos, por eso, remover deleciones que se encuentran en estas ubicaciones al inicio del análisis facilita los análisis posteriores.

Para descartar las deleciones que intersectan con los repetidos, fue utilizada una base de datos disponible en la red y creada por la herramienta RepeatMasker (56). Esta base de datos está compuesta por una tabla de coordenadas por cada tipo de región de repetidos en cada cromosoma y sus características. Para nuestros fines, solo fueron utilizados los datos que corresponden a las coordenadas y cromosomas.

Las deleciones reportadas, cuyas bases aledañas intersectaron con la base de datos, fueron removidas. Estas mismas son las bases consecutivas presentes en el genoma de referencia ubicados en ambos extremos del gap, cabe añadir que para esta tesis fueron determinadas 300. Además se usó como condición un porcentaje mínimo de intersección para descartar las deleciones. Una representación del análisis sobre una deleción se muestra en la Figura 3.5.

Asimismo, el proceso más directo que se encontró para realizar esta tarea es modificar los intervalos de las deleciones para considerar los flancos y después usar una herrameinta extra para compararlos. Sin embargo, este proceso consume muchos recursos, por una parte es computar los intervalos, y por otra parte interpretar datos de texto plano a números, por lo que se implementó la herramienta `RMFilter`. El objetivo del programa es descartar las deleciones cuyo porcentaje de sobrelape de las regiones aledañas supere un umbral establecido.

Algoritmo

Los parámetros requeridos son la cantidad de bases aledañas y el porcentaje mínimo de intersección. La implementación requiere que los datos estén separados por cromosomas. El programa primero carga la lista de coordenadas de regiones de repetidos reportadas por RepeatMasker, las cuales están presentes en el cromosoma del archivo

a analizar. Posteriormente se revisa cada una de las lecturas del archivo de deleciones; si las coordenadas aledañas a la deleción se sobrelapan debajo del umbral especificado, entonces la lectura se despliega.

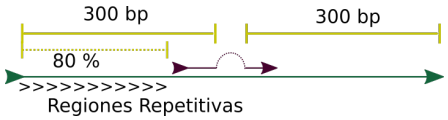


Figura 3.5: Representación del filtro por RepeatMasker
La línea morada representa la lectura que presenta la deleción. Las regiones en amarillo representan la longitud de las bases aledañas a la deleción. Los signos " > " representan una región de repetidos en el genoma de referencia (de color verde).

3.2.2.2. Sobrelape de deleciones

En este punto se tiene un archivo BED que identifica las posibles deleciones en el experimento. Aun así, se debe determinar cuáles pertenecen a una misma ubicación bajo el fin de reportar, por medio de una sola deleción, un conjunto de gaps de un mismo evento. El problema es que estos gaps no necesariamente tienen exactamente las mismas coordenadas, aunque sean parte del mismo evento. Por consiguiente, el objetivo es determinar los conjuntos que se sobrelapan entre sí. Es importante mencionar que no cualquier grado de sobrelape implica que las deleciones provienen del mismo evento.

Entonces fue necesario tener una manera de medir el grado de sobrelape dadas dos deleciones, para después agrupar las que cumplan una medida específica. Adicionalmente, fue requerido determinar el intervalo que representa este subconjunto. Las definiciones de los parámetros y modelos requeridos son desarrolladas a continuación.

Sobrelape

Para poder agrupar las deleciones primero fue definida la manera de medir el sobrelape entre dos deleciones. Entonces la definición 3.2.1 indica la forma de calcular la razón de sobrelape. Cabe anotar que el sobrelape no es una operación conmutativa, es decir, que su valor depende del orden. Este concepto es representado en la Figura 3.6, donde la razón cambia según el orden de las deleciones en la función *sob*.

Definición 3.2.1 *Sean A y B dos deleciones con sus respectivas coordenadas inicial (X) y final (Y), llamadas X_A, Y_A, X_B y Y_B. La razón de sobrelape $sob(A, B)$ es el cociente de la región intersectada de las coordenadas, entre la región de B.*

$$sob(A, B) = \frac{min(Y_A, Y_B) - max(X_A, X_B)}{Y_B - X_B} \tag{3.1}$$

Es posible definir la propiedad de respaldo (Definición: 3.2.2), y de respaldo recíproco (Definición: 3.2.3), los cuales dependen de un margen arbitrario h.

Figura 3.6: Definición de sobrelape dados los porcentajes de sobrelape
En este ejemplo se observa cómo difiere el sobrelape entre dos deleciones, dependiendo del orden. La intersección de ambas deleciones es delimitada con la línea punteada. Usando el 50 % como referencia, se puede ver que $sob(A, B) < 0.5$ mientras que $sob(A, B) > 0.5$

Definición 3.2.2 *Dadas dos deleciones A, B y un umbral mínimo h tal que $h \in [0, 1]$. Se dice que A respalda a B si y solo si $sob(A, B) \geq h$, lo cual se denota como $A \to B$.*

Definición 3.2.3 *Dadas dos deleciones A, B y un umbral mínimo h tal que $h = [0, 1]$. Se dice que existe un respaldo recíproco entre A y B si y solo si $A \to B$ y $B \to A$, lo cual se denota como $A \leftrightarrow B$, equivalente a $B \leftrightarrow A$.*

Es necesario subrayar que un respaldo recíproco no es una propiedad transitiva, esto quiere decir que aunque se den tres deleciones, esto es, $A \leftrightarrow B \leftrightarrow C$, no es suficiente para determinar si $A \leftrightarrow C$, tal y como se nota en la Figura 3.7.

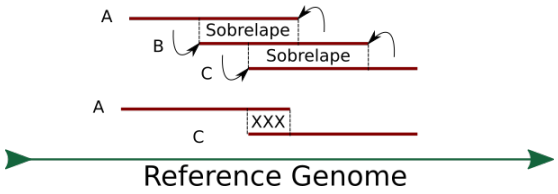


Figura 3.7: Ejemplo de recurrencia entre lecturas distintas
En este caso se logra observar cómo es que el respaldo recíproco no tiene propiedad transitiva. Por lo cual es necesario revisar todos los posibles candidatos de sobrelape por cada una de las lecturas. En el primer caso, A y B tienen un respaldo recíproco, así como B y C. En el segundo caso, A y C no tienen un respaldo recíproco, porque el intervalo que comparten es muy reducido.

Esto permite cambiar el análisis en función de qué tan estricto es requerido el grado de sobrelape. En esta tesis se utilizó $h = 0.5$, puesto que esta es la razón mínima que incluye variaciones de un nucleótido en deleciones de tamaño 2.

Conjunto de sobrelapes

La estructura que se genera a partir de las relaciones de respaldo equivale a un grafo en el que estas equivalen a las aristas y las deleciones a los nodos.

Definición 3.2.4 *Sean A y B cualquier deleción en el conjunto de datos llamado V, entonces $G = (V, E)$ es el grafo tal que los nodos equivalen a las deleciones y las aristas*

se definen de la siguiente manera:

$$E = \{(A, B) \forall A, B \in V \mid A \leftrightarrow B\} \tag{3.2}$$

El problema que se planteó en este punto fue identificar los subconjuntos de nodos, donde todas las deleciones se encuentran recíprocamente respaldadas. En lenguaje matemático, el planteamiento es identificar los subgrafos inducidos completamente conexos, los cuales son llamados cliques.

Se debe notar que todo subgrafo inducido de un clique es igual a un clique. Por consiguiente, las deleciones en los cliques maximales son las que ubican un mismo evento; los cliques maximales, por definición, no pueden estar estrictamente contenidos en otros. Dicho enfoque es similar al que utiliza `Delly` (52).

Observación Sean G' y G'' distintos cliques maximales de G. Entonces $G' \not\subset G''$ ni $G'' \not\subset G'$. Sin embargo, es posible que $G' \cap G'' \neq \emptyset$

Pueden existir cliques maximales que se enciman, no obstante, no están contenidos entre sí, tal como lo indica la observación anterior; ello implica que las deleciones pueden presentar casos como se muestra en la Figura 3.8. De esta forma, filtros posteriores no descartaron eventos representativos dentro de la misma región.

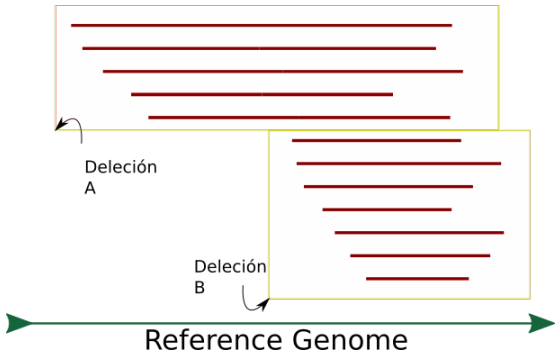


Figura 3.8: Reporte de deleciones encimadas
Cada una de las líneas representa una deleción, ordenadas por la primera coordenada. El programa implementado ubica las deleciones que son recíprocas entre sí de manera separada, representado por los recuadros. Por lo tanto, la información de las deleciones, las cuales están sobrelapadas entre sí sin ser recíprocas, es conservada.

Representación de conjuntos

Por último, queda el problema de cómo representar un conjunto por medio de un solo registro. Un conjunto de deleciones es representado por la región cubierta por todas. Por medio de un clique maximal G' se obtiene un par de coordenadas X, Y por medio de la ecuación 3.3. Para el cálculo anterior se considera que X y Y son los conjuntos de las coordenadas iniciales y finales de las deleciones contenidas en los grafos respectivamente.

$$X, Y = (max(X_{G'}), min(Y_{G'})) \tag{3.3}$$

Algoritmo

Para poder llevar a cabo este análisis completo se implementó una herramienta llamada `Del_Overlap`. El programa lee una lista de intervalos ordenados y agrupados por cromosoma, y posteriormente, se ejecutó el algoritmo por cada una de las líneas.

Primero la línea actual es marcada como la deleción principal y es guardada en la memoria intermedia. En esta última, son almacenados los siguientes registros, siempre y cuando se sobrelapen con la lectura inicial. Una vez que ya no es posible que exista otra lectura que sobrelape se deja de leer la entrada. Este proceso está ejemplificado en la Figura 3.9.

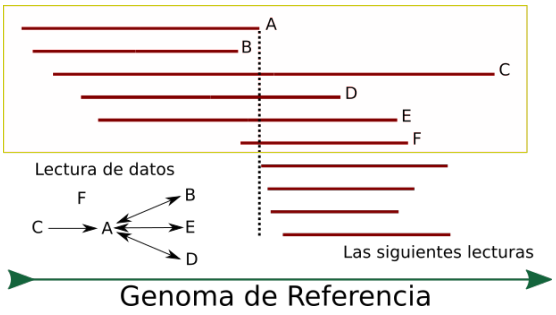


Figura 3.9: Representación de la obtención de los grafos dentro del algoritmo de detección
El rectángulo amarillo representa las deleciones que son evaluadas para añadir a la memoria intermedia, delimitado por la última coordenada de la primera lectura (línea punteada). Empero, solo se añaden aquellas que tengan un respaldo bidireccional con "A". Las otras aristas se calculan después.
La deleción "A" es la referencia para revisar sobrelapes, por lo que las aristas entre los nodos del complemento no están representados.

Una vez que se guardaron los datos en la memoria interna, entre todos sus elementos son calculados los respaldos. Como resultado se obtiene un grafo en el cual se buscan los cliques maximales, que contengan la deleción inicial. Aquellos cliques maximales, que no estén contenidos en alguno que se haya reportado previamente y superen un mínimo de nodos, son reportados.

Como resultado se obtiene otro archivo en formato BED, el cual contiene las representaciones de cada uno de los cliques maximales. En él se reportan las coordenadas y la cantidad de nodos de los conjuntos de deleciones calculados, como se representa en la Figura 3.10.

Fueron analizados intervalos de diversos tamaños, desde uno hasta 10k nucleótidos aproximadamente; la implementación soportó una cantidad de 100k deleciones en el mismo experimento durante el proceso de comparación.

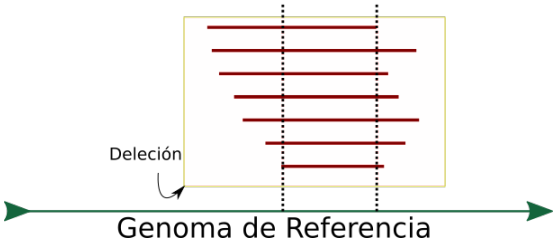


Figura 3.10: Forma de representación de deleciones sobrelapadas
El grupo de deleciones que se sobrelapan entre sí (en rojo) tienen una representación con las coordenadas marcadas con la línea punteada, las cuales corresponden a las coordenadas máxima inicial y mínima final del grupo.

3.2.2.3. Filtro somático

Hasta este punto fueron ubicadas las deleciones por cada experimento, el siguiente paso es determinar cuáles son de origen somático; las que no sean de origen somático son entonces germinales. La característica principal de las germinales es que todos los genomas celulares del individuo las contienen, debido a que están presentes desde las primeras fases de la gestación. Una deleción germinal puede estar en uno o dos cromosomas, porque somos organismos diploides.

Como consecuencia, un conjunto de lecturas que mapean a una región con deleción germinal presentarán el gap en un 50 % o 100 % de ellas. En contraste, un conjunto de lecturas que mapean a una región con deleción somática presentarán el gap en un porcentaje distinto al 0 %. El objetivo del filtro somático es remover las deleciones cuya región presente gaps en el 50 % o 100 % de ellas. Cabe mencionar que parte de las deleciones somáticas se pierden, como se muestra en la Figura 3.11.

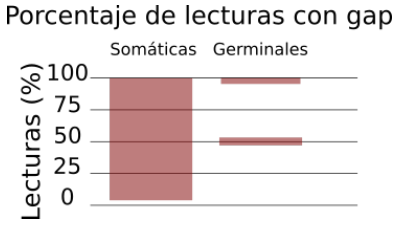


Figura 3.11: Rangos de porcentajes de lecturas con gaps, dependiendo del origen
Los rangos que ejemplifican el porcentaje de lecturas que presentan una deleción son las regiones sombreadas de rojo. Las deleciones somáticas pueden estar representadas en un rango estrictamente mayor a 0 %. De manera análoga, las germinales se presentan como gap en el 50 % o 100 % de las lecturas.

Debido a que los mapeos no están distribuidos uniformemente, se utilizó la profundidad para aproximar los orígenes somáticos y se revisó la profundidad media que existe en las regiones de la deleción y sus flancos. La razón que existe entre ambos tipos de regiones, es complemento de la fracción de lecturas. Entonces, las deleciones

germinales corresponden a una razón de profundidad de 0.0 o 0.50, con respecto a los flancos, las cuales fueron descartadas. Sin embargo, fue necesario considerar que en el caso del sexo masculino, los cromosomas sexuales son disntintos. Por lo tanto, la razón de profundidad germinal en los cromosomas "X" y "Y" en este caso solo es 0.0.

La fracción de profundidad fue definida como el cociente de la profundidad media dentro de las coordenadas, entre la profundidad de sus respectivos flancos (fórmula 3.4). En esa medida, el problema a resolver en este paso es obtener las profundidades medias en ambas regiones por cada deleción reportada. Con lo anterior se procede a descartar las deleciones, que tengan una fracción de profundidad igual a 0.0 o 0.5.

$$r = \frac{\bar{C}_{del}}{\bar{C}_{flanco}} \tag{3.4}$$

Implementación

En primera medida fue necesario realizar el cálculo de profundidad por posición en el genoma de referencia. Así pues, la cuenta debe contemplar que existen regiones que no aportan a la profundidad dentro del intervalo de una lectura, marcadas por "N" y "D" del CIGAR. La tarea fue realizada por la herramienta `genomecov` de `bedtools` (51), que calcula la profundidad en cada base del genoma. Para esto fueron requeridas las lecturas contenidas en el archivo de alineamiento en formato BAM.

En el siguiente paso fue creada una tabla para computar las fracciones de profundidad. Esta contiene los intervalos de deleción y sus flancos, junto con la suma de profundidades en dichas regiones. Los flancos fueron definidos como las regiones que distan dentro de los 20 nucleótidos fuera de los extremos de la deleción. A fin de realizar este paso fue desarrollada una aplicación llamada `Coverage_count`. Después se aplicó el filtro, llamado `Allelic_Filter`, que calcula la profundidad media en cada una de las regiones. Con este valor se puede calcular la fracción de profundidades y con ello filtrar las que son iguales a 0.0, 0.5 o 0.1, como se muestra en la Figura 3.12.

Dado que la profundidad en una región no es regular, los promedios calculados dentro y fuera de la deleción no son valores enteros. Así pues, se tendrá una incertidumbre en cuanto a la fracción real que se calcule después. Se considera una lectura extra o faltante dentro de la deleción, con una profundidad promedio de 40X, resultando en una incertidumbre de ± 0.025, es decir, fueron descartadas las lecturas cuya fracción entrara dentro del intervalo de incertidumbre alrededor de las fracciones 0.0, 0.5 y 0.1.

3.2.3. Deleciones compartidas entre individuos

Para determinar si las deleciones encontradas se presentan en una región específica del ADN, tienen que ser frecuentes en más de un individuo. Entre mayores observaciones existan en distintos tejidos o individuos, más evidencia apunta a una explicación más general en el humano. El objetivo en esta sección fue obtener las deleciones que son recurrentes en todos los experimentos.

Para encontrar las coincidencias de ubicaciones entre experimentos fue utilizada la herramienta `Intersect` de `bedtools` (51); este programa revisa un archivo BED de

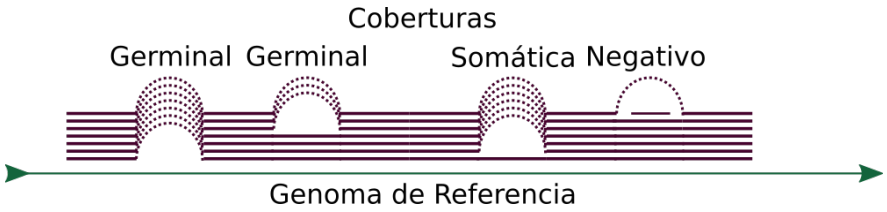


Figura 3.12: Representación de las posibles deleciones detectadas
La profundidad se representa con líneas moradas y las deleciones con los arcos punteados. Las primeras dos deleciones son las de tipo germinal que se descartan, las cuales tienen fracciones de 0.0 y 0.5 respectivamente. La siguiente deleción es somática, que en este caso, si la fracción está entre 0 y 0.5. Por último, hay una deleción que no se considera como somática puesto que hay una profundidad alta dentro de la deleción, por lo que la fracción es 1.

referencia, contra uno o más archivos de búsqueda. Fueron reportadas las deleciones que tengan respaldo recíproco con una razón mínima de 0.5, como se observa en la Figura 3.13.

Las relaciones de respaldo recíproco que son reportadas no contemplan las relaciones entre los archivos de búsqueda. Para compensar esta restricción, la intersección fue repetida usando distintos experimentos como referencia. Por lo tanto, una deleción recurrente está presente por lo menos tantas veces como el número de experimentos que fueron analizados, cuatro en el caso de esta tesis. Una mayor representación de una región, implica la existencia de deleciones que sobrelapan fuera del umbral mínimo, como se muestra en la Figura 3.14, lo cual implica que estas regiones presentan muchas alteraciones.

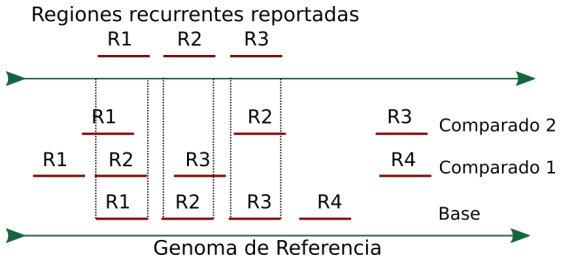


Figura 3.13: Comparación de deleciones entre experimentos
En este caso se evidencia la operación del programa `Intersect`, donde el archivo BED de "Base" contiene las deleciones que se comparan. Todas aquellas que intersecten con cualquier deleción de los grupos de "Comparados" es reportado. Hay que notar que a pesar de que "R2" de "Base" no está respaldado en todos los experimentos, sí es reportado. Pese a que "R3" de "Comparado 2" y "R4" de "Comparado 1" se sobrelapan, no son reportados.

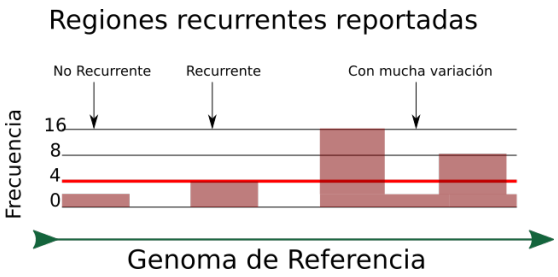


Figura 3.14: Comparación de frecuencias de deleciones comparadas
Las deleciones que se obtienen al final son producto de comparar todos los experimentos contra todos, para reportar las que son respaldadas de forma recíproca entre experimentos. En este caso se usaron cuatro experimentos, por lo que la referencia es una frecuencia de cuatro deleciones iguales reportadas, es decir, las deleciones que aparecen en todos los experimentos son los que tienen frecuencia 4. Sin embargo, aquellos que aparecen en menor frecuencia no son compartidos y los que tienen mayor frecuencia presentan mayor variabilidad.

3.3. Intersección con cáncer

Fue llevada a cabo una intersección con genes relacionados con cáncer, definidos con fundamento en el censo que existe de genes mutados en la base de datos del Genome Data Commons (GDC) (40). A partir de este se descargaron los genes relacionados con cáncer, los cuales corresponden al campo llamado "Is Cancer Gene Census". Posteriormente fueron obtenidas las coordenadas de los genes con respecto al genoma de referencia "GRCh38/hg38" por medio del portal de UCSC (30). La lista se guardó como formato BED para obtener la intersección de la base de datos con la lista de cáncer. Esta lista contiene 573 genes de la base de datos que corresponden a 5069 sitios en el genoma. La intersección fue realizada con la herramienta `intersect` de `bedtools` de manera que se conservó la información de los genes intersectados y las deleciones. Para no descartar cualquier efecto que las deleciones puedan tener fue buscado cualquier tipo de intersección en tamaño y cobertura. De igual manera, se utilizaron las regiones completas de los genes, sin excluir las regiones no codificantes.

3.4. Automatización

Las herramientas desarrolladas constan de programas necesarios para llevar a cabo todo el análisis, así como scripts que ejecutan paso a paso los segmentos descritos. Los scripts permiten que el usuario sea capaz de reemplazar los datos a analizar sin necesidad de editar cada uno de los pasos. A su vez proveen de la facilidad para poder modificar los parámetros de los procesos de detección de deleciones y filtrado.

Existen tres scripts principales, el primero lleva a cabo la descarga, el curado y el alineamiento de las lecturas crudas, lo que permite la inspección de la calidad en

cada paso. El segundo script realiza la detección de deleciones y el filtrado por cada experimento.

El último establece la comparación de deleciones entre experimentos. El proceso completo está resumido en la Figura 3.15.

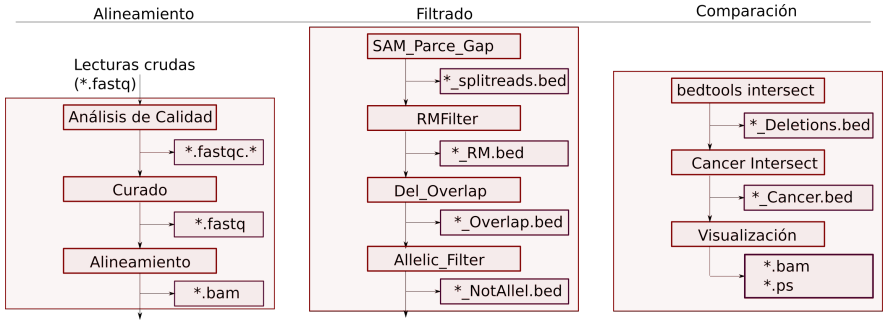


Figura 3.15: Representación del proceso automatizado
En el esquema se ven tres bloques que corresponden a cada script que automatiza el proceso. El primero lleva a cabo la descarga, curado y alineamiento de lecturas. El segundo aplica los filtros descritos dentro de un experimento. El tercero utiliza la información en distintos experimentos para comparar deleciones y visualizarlas. Dentro de los recuadros hay bloques secuenciales que indican el proceso efectuado, y los recuadros laterales representan los archivos intermedios que derivan, son los archivos generados por cada proceso y son representados por una expresión regular de los nombres utilizados.

Capítulo 4

Análisis de Resultados

Por cada experimento, se extrajeron las deleciones que fueron semejantes a otras presentes en los otros tres experimentos. Sin embargo, estas últimas no necesariamente tienen que estar respaldadas por todos los demás experimentos. Para poder confirmar esto, en las deleciones resultantes, se juntaron las que fueron detectadas en cada uno de los experimentos. Para la visualización, las deleciones resultantes fueron separadas a partir de su orden de longitud. Posteriormente, los datos fueron visualizados por medio de "igv viewer" (28). Esta herramienta permite observar el mapeo de lecturas en regiones específicas, por medio de una representación de la cobertura de las lecturas en la parte superior de cada experimento, así como la representación individual de las lecturas en la continuación de las gráficas.

Por cada lectura, la región alineada es representada con regiones grises, los cambios de bases se representan con un color que corresponde a la base presente en la lectura, los gaps se representan como líneas delgadas y las inserciones como marcas moradas con el símbolo "I".

Las regiones encontradas se muestran como marcas en rojo que se muestran en la parte superior de la interfaz de "igv viewer". Por lo tanto, en las imágenes que fueron obtenidas de dicho programa, contienen estas zonas marcadas en la parte superior.

Tener varios puntos de corte es indicio de un evento que ocurre de forma somática en distintas células. En consecuencia, los gaps en las lecturas de una deleción puede que no estén sobrelapadas completamente. De igual manera, debería apreciarse un valle en el perfil de cobertura. Un ejemplo de una deleción en una ubicación definida es la figura 4.1. Sin embargo, existe la posibilidad de que estas irregularidades en los puntos de corte sean consecuencia de mapeos erróneos o con distinta representación.

Aún es posible tener mapeos ambiguos a consecuencia de regiones donde subsecuencias pueden tener más de un posible alineamiento. Por ejemplo, una deleción dentro de una cadena de "A" repetidas, es un descubrimiento ambiguo, como se muestra en la Figura 4.2.

Es necesario añadir que en este análisis se excluyeron los resultados del cromosoma "Y", puesto que solo dos experimentos corresponden a tejidos provenientes de hombre, lo cual no permite una comparación en más de dos experimentos.

Con el fin de identificar todas las posibles ubicaciones posibles, fueron incluidas las deleciones que se encuentran en segmentos no identificadas en los cromosomas. Estos tienen identificadores especiales en el campo del cromosoma para separarlos de los demás.

Cabe mencionar que como parte del método de visualización, solo se representan las lecturas que intersectan con las deleciones. Esto permitió llevar a cabo una visualización más clara, sin embargo, crea un efecto de coberturas nulas después de las regiones de los flancos de las deleciones.

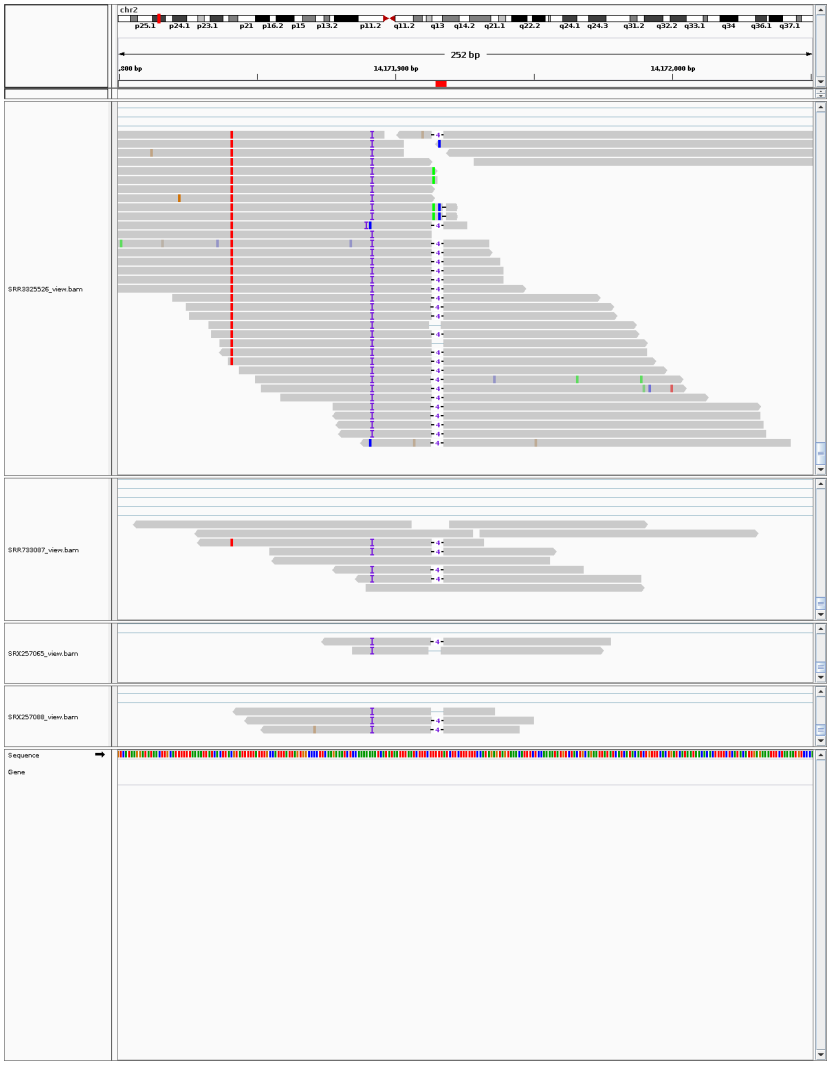


Figura 4.1: Ejemplo de una deleción en una región con mapeos aledaños muy variables. Las lecturas que cubren el área detectada con deleciones intersectadas. Cabe mencionar que la alta variación en el genotipo de las lecturas ponen en duda el mapeo de estas lecturas.

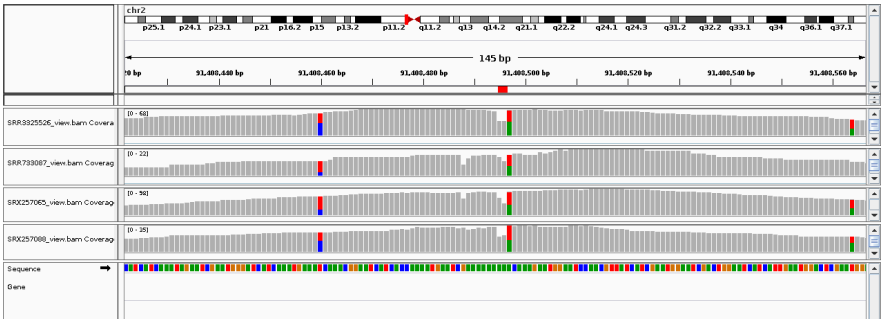


Figura 4.2: Ejemplo de una deleción en una región con ambigüedad
La serie de "A" (color verde) en el genoma de referencia indica que la deleción detectada cae en una región de bases iguales consecutivas. Por lo tanto, un mapeo con deleciones recorridas sobre dicha subsecuencia es igual de válida, sin embargo, no es posible saber la ubicación de la deleción de manera definitiva.

4.1. Frecuencia de longitudes

Se obtuvieron distintas cantidades de deleciones en cada uno de los experimentos presentados en la Tabla 4.1. Gran parte de ellas corresponden a deleciones de 100 b a 100 Kb. La frecuencia disminuye en función de la longitud a partir de la longitud de dos bases, como se muestra en la Figura 4.3. Por otra parte, en la Tabla 4.1 se muestran las deleciones de cada experimento, separados por orden de magnitud. En la última fila se representa la unión de las deleciones corroboradas en cada experimento. Por lo tanto, la cantidad total de este conjunto puede ser a lo más la suma de deleciones en cada experimento respalda.

En la figura 4.3 se muestra que parecen existir tres grupos de deleciones que se pueden separar por intervalos de longitud, las menores de 100 bases, las de 100 bases a 100 Kb y las mayores a 10Mb. Puede proponerse que esta diferencia se debe a distintos mecanismos que tienden a provocar modificaciones a distintos intervalos del genoma.

Por otro lado, aún está presente una gran cantidad de variación, por lo que esta gráfica puede cambiar si se hacen más estrictos los métodos. Las deleciones más grandes tienen mayor libertad para variar. Por esa razón, es más probable que la recurrencia de estas sea consecuencia de eventos aleatorios que coinciden en una región. En esa medida, las deleciones más cortas son las que tienen una mayor confianza de ser comprobadas experimentalmente.

4.2. Frecuencia de sobrelape

Las deleciones finales que fueron graficadas tienen que estar sobrelapadas sobre una misma región. Existen regiones en el genoma que son cubiertas por distintas deleciones. Esta cantidad es llamada frecuencia de sobrelape, la cual fue graficada por medio de

Nombre	***Número de registros con longitudes en los intervalos :***								
	$[1, 10)$	$[10, 100)$	$[100, 1K)$	$[1K, 10K)$	$[10K, 100K)$	$[100K, 1M)$	$[1M, 10M)$	$[10M, 100M)$	$[100M, 500M)$	$[1, 500M)$
SRX1660320	102	131	389	548	217	5	7	41	4	1444
SRX237626	80	69	145	253	104	4	2	23	1	681
SRX257065	91	95	233	466	187	3	6	18	2	1101
SRX257088	82	78	193	318	135	3	3	21	0	833
Unión	142	314	944	1574	639	14	18	103	7	3755

Tabla 4.1: Tabla de número de registros por experimento

Se muestran los cuatro experimentos utilizados para el análisis. Cada uno de los experimentos lo componen las lecturas contenidas en las corridas descargadas. Las deleciones de cada experimento, las cuales fueron detectadas, filtradas y compartidas en los experimentos restantes. Cada campo separa el orden de magnitud. Al final se encuentra la unión de todas las deleciones encontradas.

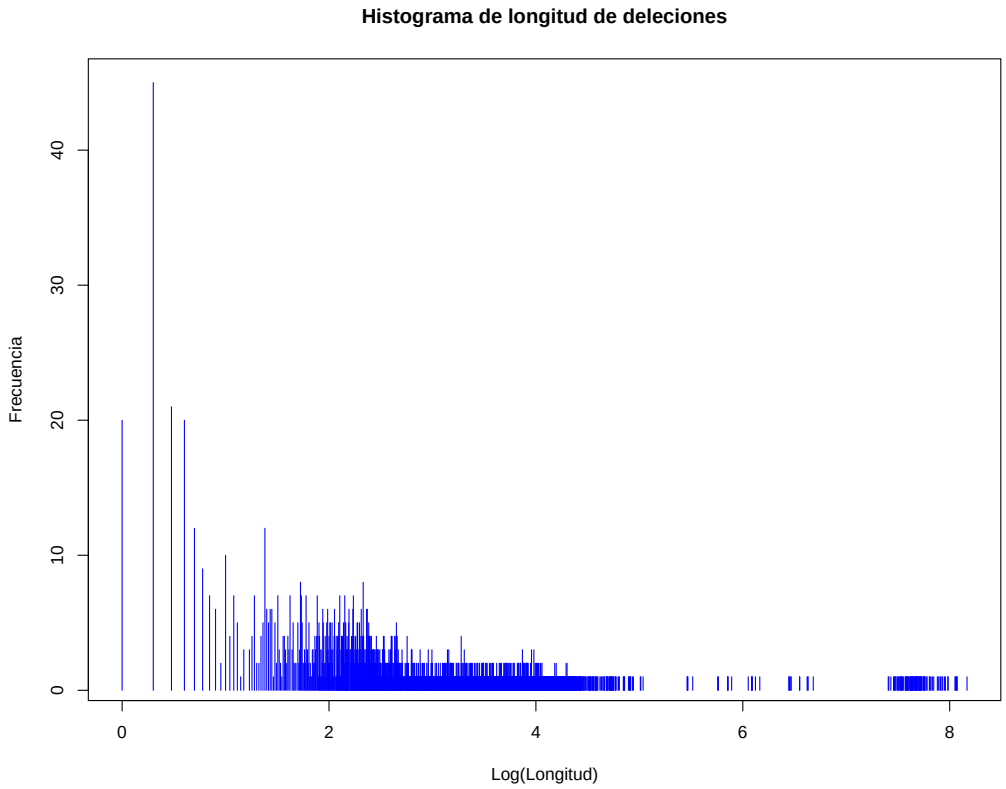


Figura 4.3: Dispersión de frecuencia de las longitudes de las deleciones compartidas entre experimentos
La gráfica presenta el número de deleciones que existen por cada uno de los valores de longitud existentes. En pos de identificar la diferencia entre las deleciones de distintos órdenes de magnitud, la escala de las longitudes se muestra logarítmica.

la herramienta `KaryoploteR` (18) y se encuentra en el apéndice de esta tesis. Se espera que la frecuencia de una deleción sea por lo menos de cuatro para indicar la recurrencia en los cuatro experimentos, y que no supere el doble de esta medida, ya que esto indica alta variabilidad y por lo tanto baja confianza de descubrimiento.

Aunado a ello, fueron representadas deleciones individuales en los mismos gráficos en el apéndice. Esta representación permite representar la información por cada par de coordenadas. De esta manera es posible comparar la forma en que sobrelapan las deleciones y la frecuencia de deleciones dentro las mismas ubicaciones.

En los distintos experimentos se observó que muchas deleciones se sobrelapan dentro de una misma región con mucha frecuencia, lo cual indica regiones de mucha variabilidad. Principalmente estas regiones se ubican en los centrómeros y subtelómeros como se ve en los cromosomas 2, 3, 6, 7, 10, 17 y 20. Sin embargo, las regiones con deleciones dispersas también se encuentran en otras partes dentro de los cromosomas, un posible

patrón se explora en la comparación con genes de cáncer.

Existen deleciones con la frecuencia esperada en gran parte de los cromosomas. De igual manera hay regiones de alta frecuencia que pueden estar compuestas por grupos de cuatro o cinco deleciones que se sobrelapan entre sí en altos porcentajes. Por esto, es posible considerar dichas regiones como la superposición de distintas deleciones con frecuencia esperada.

Las deleciones de menor longitud tienen mayor fiabilidad al ser más precisas, a consecuencia de que al ser más cortas hay menor libertad en la ubicación de las demás deleciones entre distintos experimentos. Aun así, los principales criterios para elegir el subconjunto de deleciones que se comprobarán experimentalmente dependen de que estén fuera de regiones de alta variabilidad o de secuencias repetidas.

Por un lado, estas representaciones permiten identificar regiones en las que hay mucha variación de gaps, los cuales corresponden a niveles altos de frecuencia. De igual manera, es posible identificar deleciones recurrentes con frecuencias cercanas a cuatro e incluso identificar deleciones superpuestas gracias a la representación de las deleciones de forma individual. Una visualización a otra escala de las regiones de alta frecuencia podría ayudar a descubrir otras deleciones superpuestas que parezcan regiones de alta variabilidad.

4.3. Características de las deleciones

4.3.1. Deleciones proximas a centrómeros y subtelómeros

Gran parte de las deleciones están ubicadas dentro de regiones cercanas a los centrómeros, lo cual es consistente con la variación genómica que existe en las regiones cercanas al centrómero, las cuales son susceptibles a transposiciones (55). Se puede ver en la Figura 4.7 la variación presente en la región aledaña a la deleción cercana al centrómero.

También se encuentran deleciones en las regiones subteloméricas, con mucha variabilidad entre lecturas de manera similar a los centrómeros, así concuerdan con la variación genómica observada en dichas regiones (42). La Figura 4.6 es un claro ejemplo de variación en las regiones cercanas a los telómeros.

Dicha naturaleza no es exclusiva en las regiones teloméricas o en centrómeros, pues existen otras de alta variación fuera de las mismas. Si bien la Figura 4.1 es un ejemplo de esto, no es tan común. Fueron revisados casos en los que existe una baja frecuencia de deleciones dentro de estas ubicaciones, los cuales presentan mucha variación entre las lecturas presentes, como se presenta en la Figura 4.4. Aun así, existen pocas deleciones cercanas a centrómeros que no tienen una alta variación, como en la Figura 4.8, que presenta mapeos uniformes.

En vista de que dentro de los centrómeros y subtelómeros pueden existir genes, también existen casos en los que estas regiones intersectan con las deleciones encontradas. Tal es el caso de la Figura 4.8, una deleción se encuentra cerca del centrómero, no

obstante, contiene un gen dentro de la región.

4.3.2. Regiones de alta variabilidad

Una de las condiciones experimentales es la viabilidad del diseño de cebadores para validar experimentalmente la región, sin embargo, las regiones susceptibles a transposiciones presentan un reto extra, por lo que tienen que ser descartadas. Estas regiones también se caracterizan por contener una alta frecuencia de sobrelape. Entonces las deleciones de menor cantidad de sobrelapes son los mejores candidatos para presentar como deleciones somáticas, puesto que indican que la mayoría de los gaps en cada experimento sobrelapan un alto porcentaje entre sí y uno muy bajo con los aledaños.

4.3.3. Localización de puntos de corte

Un punto a identificar aquí es el hecho de que los gaps en las lecturas no presentan coordenadas completamente uniformes; las deleciones más cortas también restringen la cantidad de posibles configuraciones de los gaps dentro de un experimento. Esta falta de uniformidad en los gaps apunta al mismo evento somático que ocurrió en distintas células hermanas como se representa en la Figura 4.9.

4.3.4. Regiones de repetidos

Se debe tomar en consideración especial las deleciones que caen dentro de regiones de bases consecutivas, por ejemplo, las deleciones de las figura 4.2. Estas deben ser descartadas como posibles variaciones, puesto que no es posible ubicar las deleciones presentes en el alineamiento de manera única. Una ambigüedad menor de la ubicación puede ser ignorada dependiendo de la extensión de la deleción, porque puede caer dentro de la variación de los puntos de corte.

4.3.5. Intersección con genes

También se observaron distintas deleciones que caen en regiones dentro de genes; esto puede ser de interés porque es más probable que tengan efectos sobre la transcripción de proteínas. Estos eventos en específico son escasos y los que se presentan caen en regiones intrónicas, tal como se ve en la Figura 4.10. Para obtener información más precisa y detallada será necesario hacer una comparación completa con la anotación actual de genes.

4.3.6. Deleciones resultantes

Las deleciones que se comprobarán experimentalmente son aquellas que contienen una clara presencia en todos los individuos, no tienen ambigüedad de mapeo y caen en

regiones de poca variabilidad entre las lecturas. Un ejemplo se muestra en la Figura 4.11. Se debe dar prioridad a las deleciones de menor longitud, en la medida en que las de mayor longitud tienen menor fiabilidad explicada en los puntos anteriores. No obstante, es posible cambiar los criterios para deleciones mayores a 10 kilobases en análisis posteriores.

4.4. Intersección con los genes en Cáncer

Como se ha indicado en líneas anteriores, una de las enfermedades que tiene relación con las deleciones somáticas es el cáncer, por lo que se propuso comparar las ubicaciones de los genes relacionados con esta enfermedad y las deleciones encontradas. Solamente se buscaron intersecciones con los genes más representados, obtenidos de un censo que existe de genes mutados en la base de datos del Genome Data Commons (GDC) (40).

Todas estas regiones genómicas se marcaron como referencias dentro de las imágenes utilizadas para representar las deleciones. De este modo, fue posible comparar la frecuencia y los puntos de corte de las deleciones con las coordenadas de los genes asociado con esta enfermedad. Con esto fue posible identificar que existen regiones de alta variabilidad cercanas a unos genes de cáncer.

Fueron obtenidas 108 regiones intersectadas distribuidas en los cromosomas de 2, 4, 5, 6, 7, 9, 10, 11 y 17; 104 regiones pertenecen al rango de longitud de 10 megabases a 500 megabases, por lo que es imposible que no tengan intersecciones en alguna otra posición. El complemento consta de cuatro deleciones menores a 10 bases, ubicadas en el cromosoma 17.

En el cromosoma 17, todas las mutaciones menores a 10 bases intersectas con SEPT9 (1), que está relacionado con el control de las fases del ciclo celular, en particular con la separación celular durante la fisión. Adicionalmente, dicho gen se ha propuesto como gen supresor de cáncer en ovarios. La deleción está ubicada en una región de “TGG” repetidos, por lo que no se puede asegurar que todos los gaps pertenecen a la misma región, sino que provengan de una misma causa.

Aunque las deleciones en sí parecen no afectar a los genes de cáncer de manera directa. Es posible que tanto las deleciones como las regiones de alta variabilidad tengan algún efecto indirecto en los genes de cáncer.

Figura 4.4: Ejemplo de una deleción de baja frecuencia pero que presenta lecturas variables

En este caso se tiene una deleción que parece ser adecuada para comprobar con base en la frecuencia de deleción. No obstante, la variabilidad presente en las lecturas de la región no lo permite.

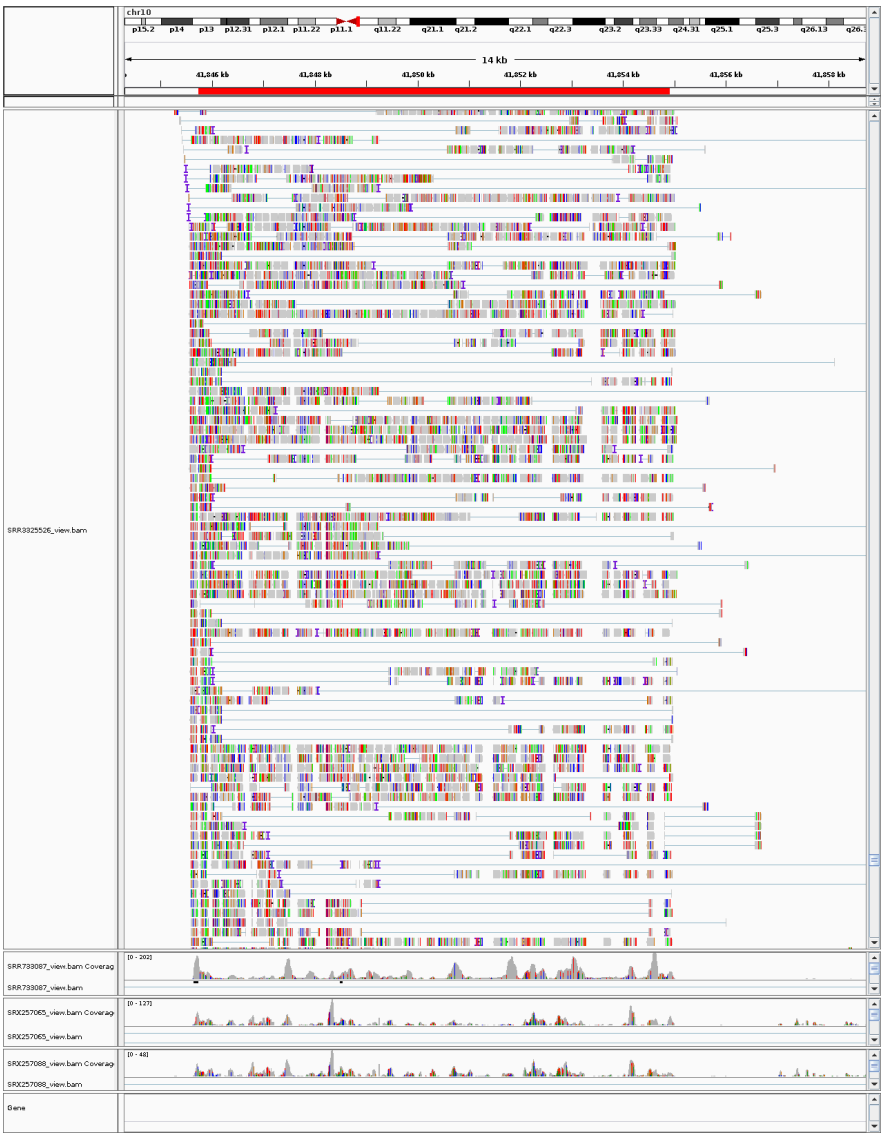


Figura 4.5: Ejemplo de una deleción en una región en el centrómero con alta variación en las lecturas aledañas

En esta figura se encuentran las lecturas mapeadas en una región cercana al centrómero. En ella se ven las variaciones estructurales con respecto al genoma de referencia y entre sí. Estas deleciones probablemente son consecuencia de las transposiciones. El intervalo marcado en rojo en la parte superior corresponde a las deleciones encimadas detectadas por las herramientas utilizadas.

Figura 4.6: Ejemplo de una deleción en una región en el telómero con alta variación en las lecturas aledañas

En este ejemplo se ven las lecturas mapeadas en una región cercana al telómero. En ella se ven las variaciones de mapeo con respecto al genoma de referencia y entre sí.

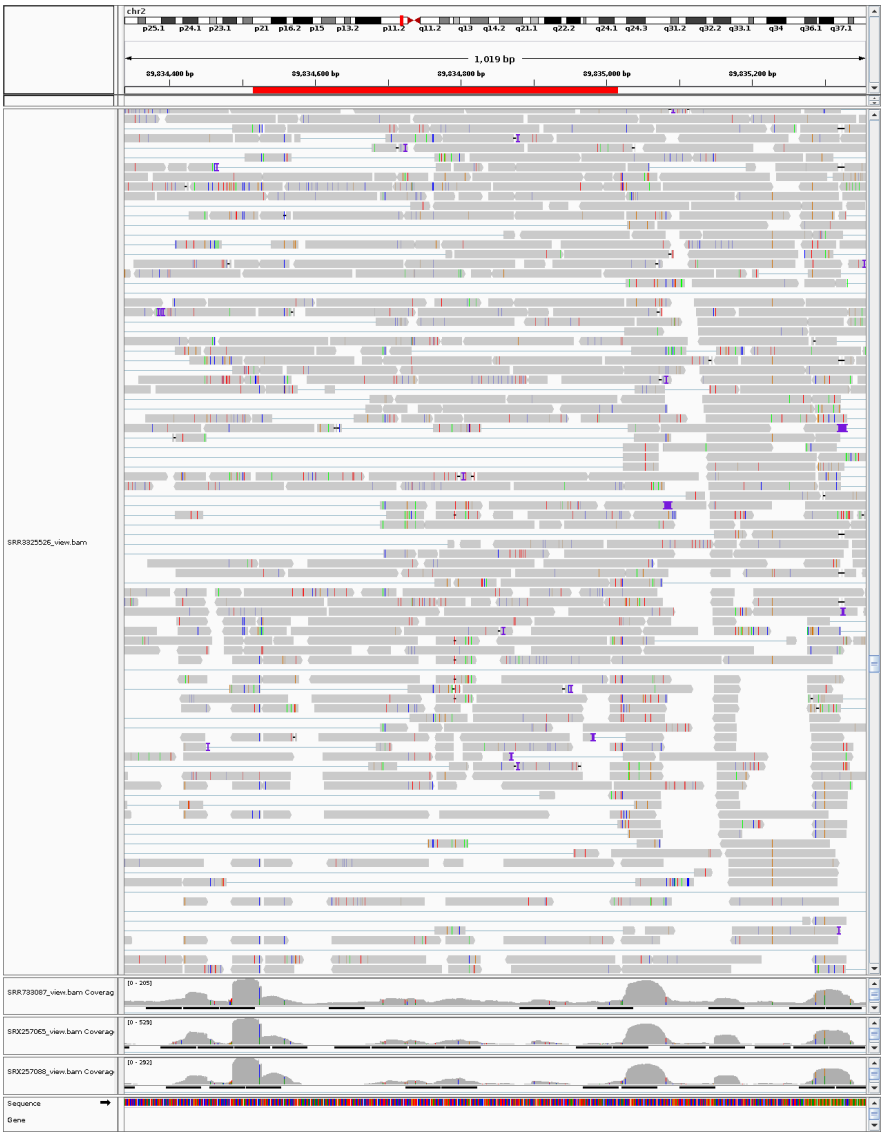


Figura 4.7: Ejemplo de una deleción de frecuencia aceptable, pero con lecturas aledañas con alta variabilidad

En este caso se observa que la deleción es recurrente entre experimentos, pese a ello, la variación que se observa en las lecturas es muy alta, lo cual impide que pueda ser comprobada experimentalmente.

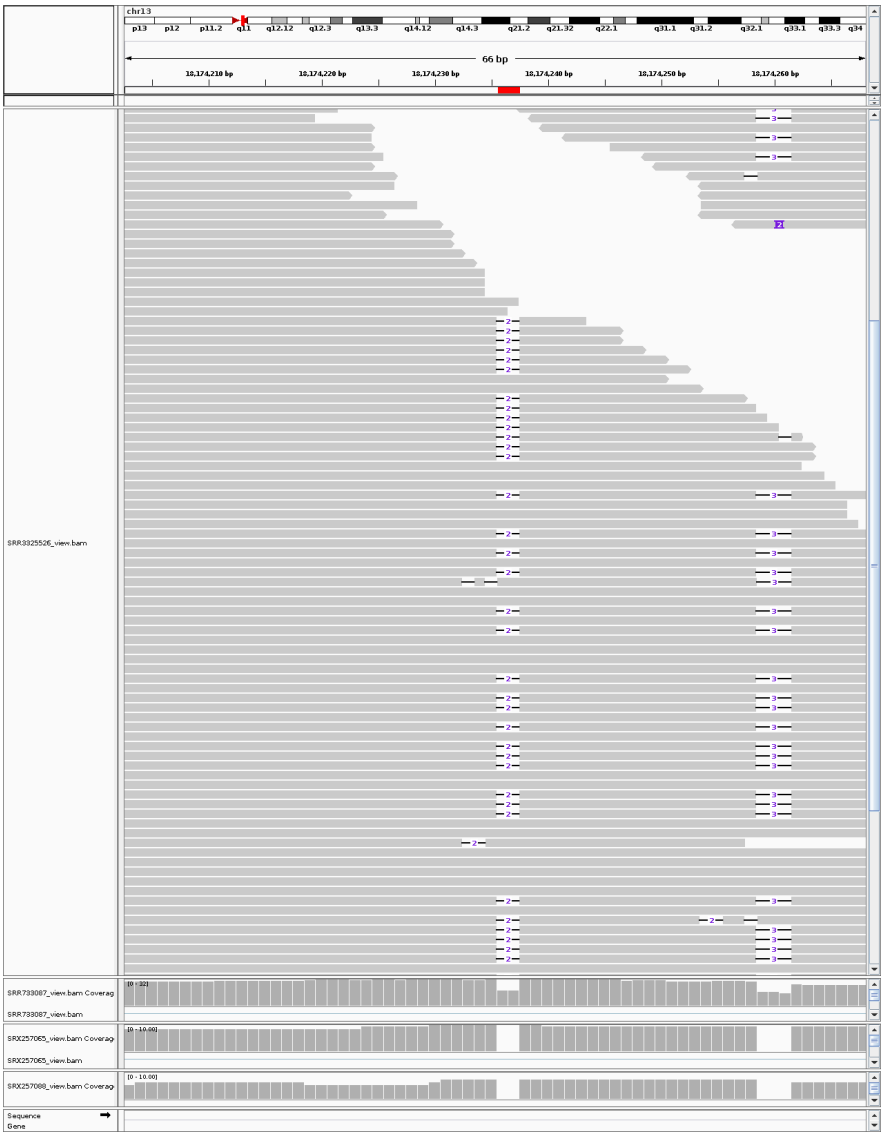


Figura 4.8: Ejemplo de una deleción en una región en el centrómero con poca variación en las lecturas aledañas

En este ejemplo se observan las lecturas mapeadas en una región cercana al centrómero, las cuales pueden ser significativas al ser consistentes en los cuatro experimentos. Sin embargo, aún tiene la probabilidad de ser una deleción de origen somático. Nosotros estamos buscando eventos que sobrelapen ya que estos tienen baja probabilidad de ser aleatorios.

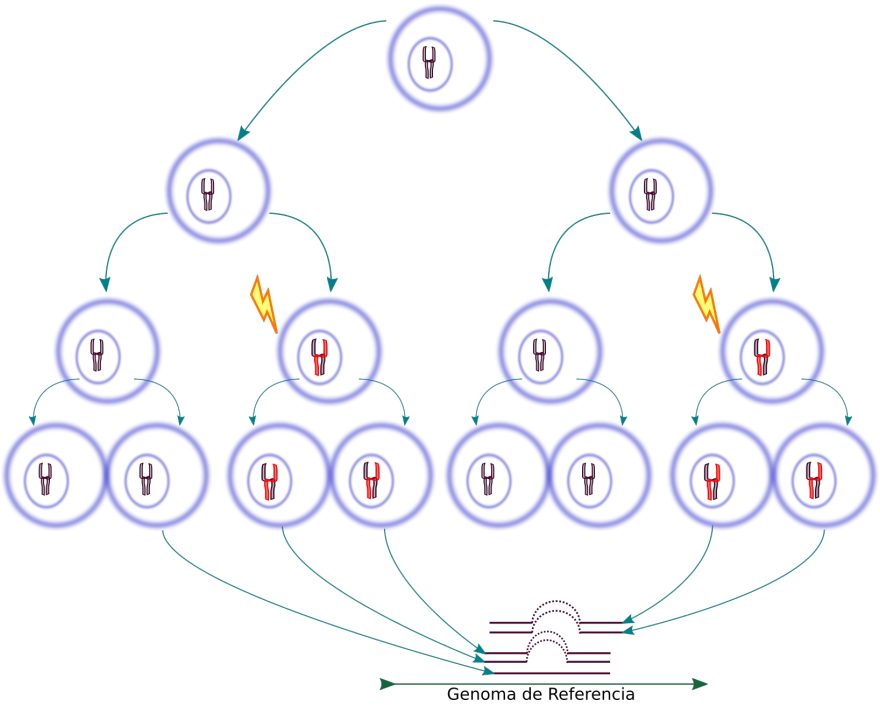


Figura 4.9: Uniformidad de eventos somáticos
La imagen muestra cómo una deleción en la misma ubicación puede tener pequeñas variaciones al afectar células hermanas. Este evento carece de uniformidad dado que contiene la información de distintas células que pasaron por el mismo efecto. Si ambos eventos fueran aleatorios, la probabilidad de conincidencia sería baja, de lo contrario, podría apuntar en la dirección de un mecanismo enfocado a una ubicación en ADN.

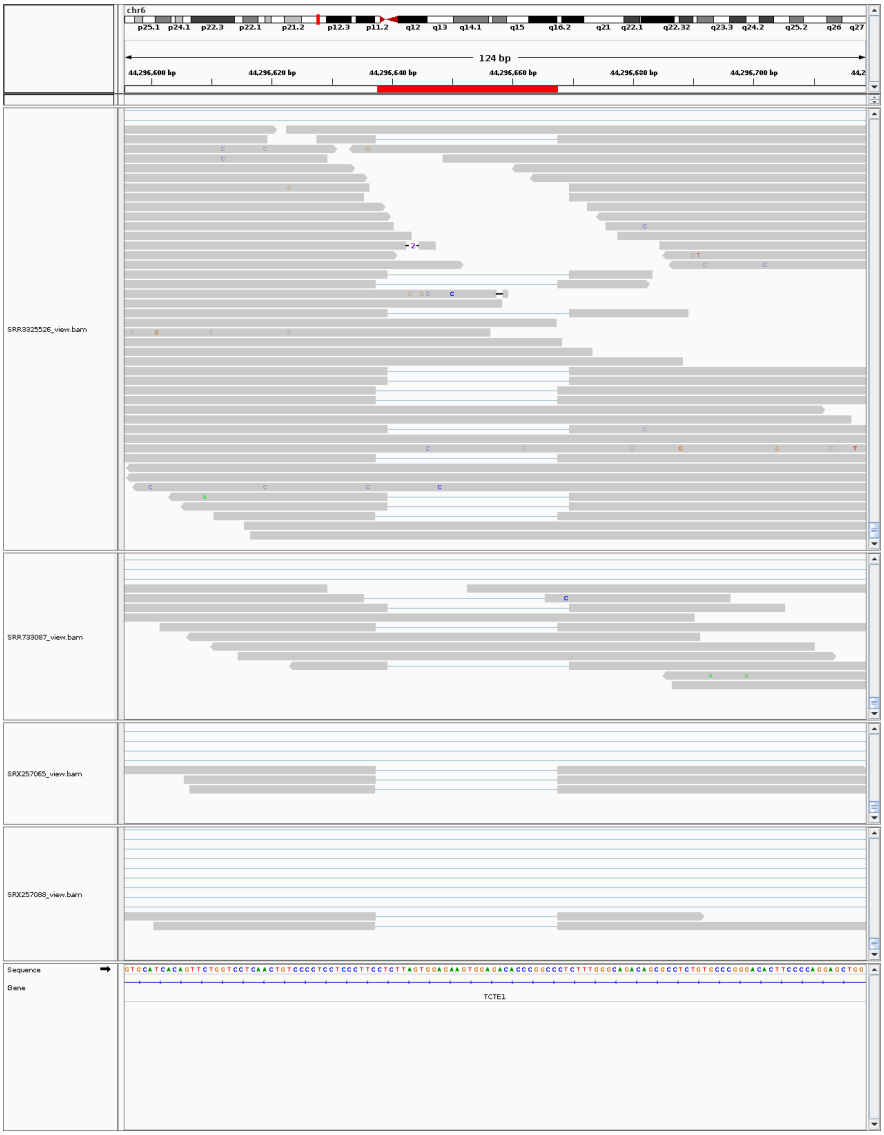


Figura 4.10: Ejemplo de una deleción intersectando un gen
En esta imagen la deleción intersecta con el gen PDE9A. La deleción tiene diferencias entre experimentos, pero puede ser consecuencia de su origen somático.

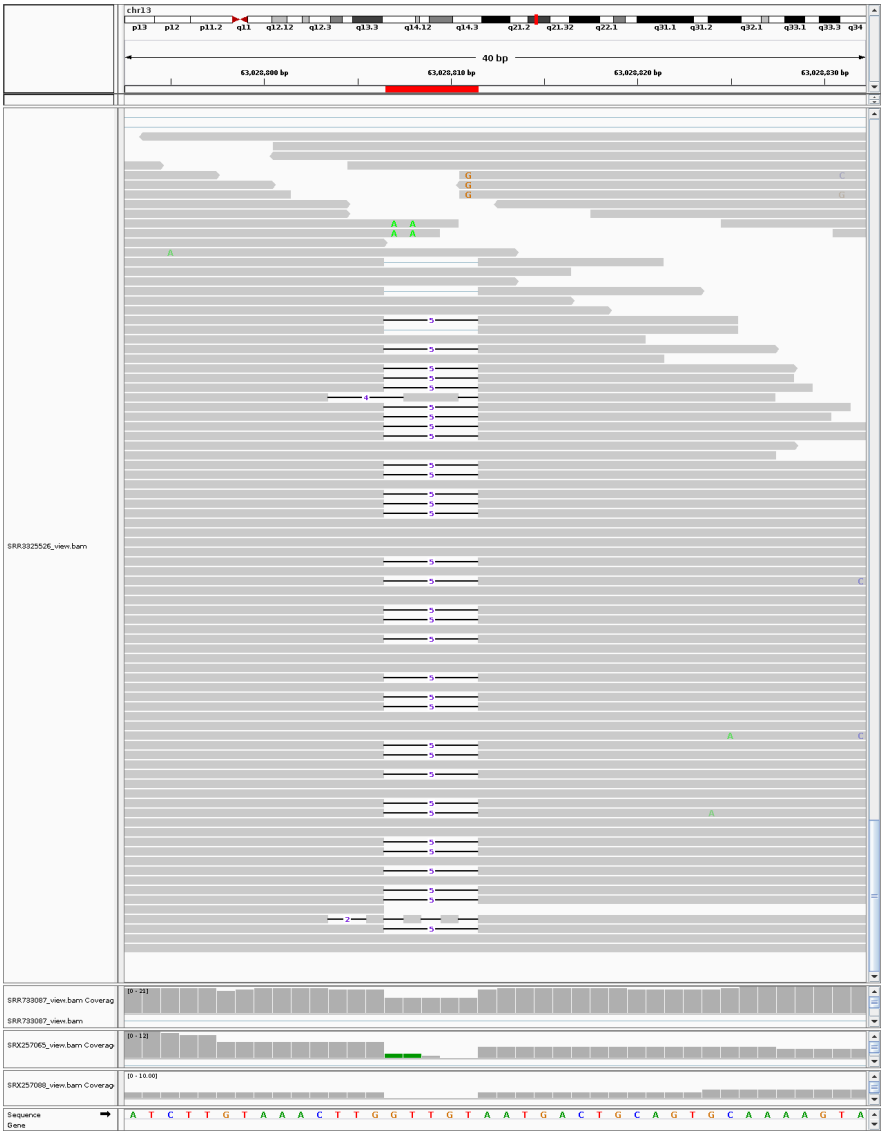


Figura 4.11: Ejemplo de una deleción con las características requeridas
Esta deleción muestra los aspectos de las deleciones que se están buscando, el tamaño, la fracción de profundidad, la poca ambigüedad de mapeo, la región aledaña de cobertura uniforme y variación reducida indican una deleción somática.

Capítulo 5

Conclusiones

A partir de este proyecto, fueron obtenidas las herramientas informáticas necesarias para poder llevar a cabo la identificación de deleciones somáticas en cualquier experimento de secuenciación, de estrategia WGS. Estos programas son de libre acceso y se pueden encontrar en `https://github.com/Bloodfield/Plasticidad_INB`. De igual manera, fue posible identificar posibles candidatos de deleciones somáticas que podrán comprobarse experimentalmente en un futuro. De forma adicional, se llevó a cabo una comparación entre la ubicación de las deleciones y los genes relacionados a cáncer. El estudio sugiere que las deleciones son del tipo pasajeras, esto quiere decir que no inducen el desarrollo de esta patología.

Fue observado que aún es requerido ajustar los parámetros de descubrimiento para obtener la mayor cantidad de candidatos a deleciones somáticas compartidas. Las herramientas desarrolladas pueden integrar más análisis, filtros y modelos que permitan refinar la detección de deleciones, así como añadir comparaciones similares a la realizada con los genes relacionados con el cáncer. Es oportuno explicar los detalles asociados con los puntos y problemas adicionales que pueden ser desarrollados a futuro.

5.1. Trabajo a futuro

Las principales vertientes del trabajo a futuro son tres; la primera corresponde a corroborar experimentalmente la presencia de deleciones en más individuos. Con esto es posible explorar nuevas hipótesis relacionadas con deleciones compartidas entre individuos.

La segunda vía corresponde a refinar las herramientas que se tienen actualmente, así como integrar unas nuevas. Es importante subrayar que actualmente las deleciones de baja frecuencia en los histogramas de frecuencia de sobrelape son las de mayor confianza, debido a que su baja frecuencia es consecuencia de la estabilidad de la región. Después de descartar el complemento, el número de deleciones que podrán ser comprobadas posteriormente es bastante menor a 8252, que se tienen actualmente.

Por último, es posible extender la comparación con la base de datos de cáncer, no

existe una intersección clara de ambos conjuntos, a excepción de las deleciones que son tan grandes que no existe probabilidad de que no intersecte con alguno de estos genes. Empero, esto solo descarta un efecto directo sobre los genes, mas no uno indirecto relacionado con cáncer. Incluso, es posible explorar otras hipótesis sobre el efecto de las deleciones sobre otras condiciones.

5.1.1. Deleciones somáticas

Antes de la comprobación experimental es necesario llevar a cabo un análisis estadístico para obtener un valor de confianza de descubrimiento. Esto implica comparar con un escenario en el que se buscan deleciones recurrentes en experimentos donde estas mismas se encuentren distribuidas de forma aleatoria. Posteriormente, la comprobación experimental consistirá en buscar estas mismas deleciones en más individuos. Para esto se requiere diseñar cebadores que se anclen a cada extremo de la región que presenta la deleción. Estos servirán para secuenciar estas ubicaciones en tejidos de distintos individuos y verificar que en efecto los gaps estén presentes.

En el caso de las deleciones de mayor extensión, fueron encontradas lecturas que mapean regiones entre los puntos de corte, lo cual puede indicar que la información genética no fue eliminada por la deleción. Se ha especulado la posibilidad de que estos fragmentos se conserven como eccDNA, dado que la información se mantiene presente pero forma parte del cromosoma. La circularización evitaría que la información genética contenida en los segmentos cortados fuera degradada por el lisosoma. No obstante, no es la única posible hipótesis, las regiones eliminadas pueden ser protegidas por otro mecanismo de síntesis de ADN, o bien, ser consecuencia de una traslocación genómica.

Sería interesante identificar posibles relaciones de las deleciones en regiones anotadas del genoma. Para ello es necesario tener una referencia de las regiones anotadas para luego hacer la intersección con las deleciones encontradas. La hipótesis en este experimento se limitó a realizar el análisis en distintos experimentos de tejido sano en distintos individuos. Ahora bien, los eventos somáticos podrían ser específicos a un tipo de tejido; tiene sentido llevar a cabo la misma comparación usando datos de secuenciación el mismo tipo de tejido entre distintos individuos. De igual manera sería interesante realizar la comparación en distintos tejidos del mismo individuo, que en principio permitirá identificar si hay eventos somáticos que ocurran de manera generalizada en un individuo, o como parte normal del desarrollo de los tejidos adultos. Las herramientas permiten modificar los parámetros de manera que se pueden hacer análisis con distinto nivel de rigor. Esto cambia la sensibilidad para detectar eventos de baja representación.

5.1.2. Software

El principal problema que se presenta con los filtros es el modelo que se utiliza para evaluar las deleciones, cuyo origen es uno de los cromosomas germinales. El filtro solo considera la cobertura local, a pesar de eso, la variación de la cobertura varía en

función de la longitud de la deleción. Por tal motivo, una solución es tener parámetros dinámicos en función de la longitud de las deleciones.

La variación de la cobertura implica también que la incertidumbre de las medidas crezca en función de la longitud. Otra propuesta es identificar las deleciones germinales con base en los demás genotipos que presenta la lectura que contienen los gaps. Si las lecturas se pueden separar en dos grupos que corresponden a cada cromosoma por medio de los genotipos, entonces será posible a la vez separar las deleciones somáticas con un criterio de cobertura más simple.

Un dato de utilidad son los puntos de corte, los cuales pueden no coincidir entre gaps. Es posible utilizar medidas estadísticas para reportar la variación que tienen las coordenadas, dentro de cada una de las deleciones. Esta información puede corroborar algún tipo de mecanismo o utilizarse como criterio de selección para su validación. El agrupamiento de las lecturas por `Del_Overlap` tiene desventajas en cuanto al procesamiento. La identificación de cliques maximales es un problema NP completo y el algoritmo implementado se puede mejorar con las heurísticas presentadas en el algoritmo Bron-Kerbosch (4).

Un caso especial a considerar son las regiones de baja cobertura, que presentan una cantidad muy baja de lecturas tanto en los flancos de corte como dentro de los gaps. Estos casos tienen que ser identificados de alguna manera, porque tienen una alta probabilidad de no ser comprobados de manera experimental.

De manera adicional, hace falta estudiar el patrón de calidad de mapeo en las lecturas que componen las deleciones encontradas. Esta evaluación puede ser un criterio extra para descartar deleciones con base en el mapeo de las lecturas cercanas. Esto significa que se puede discriminar aquellos gaps que presenten ambigüedad debido a la calidad de cada base mapeada.

Las regiones con mucha variación tales como las que se presentan en los telómeros o centrómeros tienen que ser descartadas. Aun si fueron identificadas correctamente, la corroboración experimental es más complicada si existe una taza alta de traslocaciones. Sería posible llevar a cabo una comparación cuantitativa de las deleciones detectadas por herramientas similares. Esto tiene como fin comparar las características de las deleciones detectadas por distintos métodos.

5.1.3. Deleciones somáticas y su relación con otras condiciones

En este trabajo se presentó una comparación reducida a utilizar solo genes relacionados con cáncer. También se podrían comparar las deleciones que se han encontrado recurrentes en distintos experimentos en cáncer. De igual manera, es posible enfocar los estudios al mismo tipo de tejido.

Es posible que las regiones de alta variabilidad tengan efectos sobre los genes relacionados con cáncer. Para comprobar lo anterior, es necesario ubicar todas las regiones variables, dado que en este experimento es posible que estas se hayan descartado con los filtros utilizados. En caso de que exista una consistencia de cercanía entre ambos conjuntos, lo siguiente sería buscar alguna explicación a dicho patrón.

Además del cáncer, se podrían buscar mutaciones relacionadas a otras condiciones de interés. Una de estas puede ser la progeria, en tanto que está muy relacionada con el envejecimiento. De igual manera se podrían buscar bases de datos en las que se hayan analizado mutaciones comunes en enfermedades características del envejecimiento.

5.2. Comparación extra con VarScan2

Para remarcar la motivación de este proyecto, se llevó a cabo la misma detección de deleciones con parámetros análogos utilizando una herramienta que se utliza ampliamente para la detección de polimorfismos, la cual es `VarScan2` (35). Para la detección de variaciones somáticas, esta herramienta utiliza más de un experimento de secuenciación, ya que uno tiene que corresponder a una muestra blanco que permite comparar deleciones contra otros muestra. Sin embargo, en este caso, solo se realizó el filtro de variantes con la suficiente sensibilidad para ubicar deleciones con baja representación. De manera adicional, se removieron aquellas que intersectaban con la base de datos de RepeatMasker (56). Por útlimo, fue aplicada una heurística para remover deleciones germinales. Este filtro consistió en remover variaciones con frecuencias de aparición en lecturas de 50 % y 100 %. En este análisis aledaño, fueron comparadas las deleciones presentes en los mismos cuatro experimentos de secuenciación. Al final fueron obtenidas 26 deleciones que del tamaño de 1 base en 11 cromosomas distintos y 4 regiones cromosómicas sin ubicación.

De primera instancia podemos notar la gran diferencia en el número de deleciones reportadas. Es cierto que las reportadas por `VarScan2` tienen un alto grado de confianza de ser variaciones somáticas en vez de artificios por parte de alguno de los métodos de secuenciación, alineamiento o procesamiento. Sin embargo, nosotros queremos ubicar todas aquellas deleciones somáticas que pueden parecer errores sistemáticos, pero aún tienen una pequeña probabilidad de no serlo. Aúnque el método propuesto en este trabajo no es muy limpio, evita descartar deleciones que pueden ser descartadas por estos filtros, y a la vez aplicar filtros que permitan remover aquellas que son germinales.

Otro punto a identificar es la falta de deleciones de longitudes mayores a una base. Esto puede ser consecuencia de la inexistencia de las mismas, sin embargo hay que considerar que el método de `VarScan2` analiza variaciones base por base entre las lecturas que coinciden en la misma posición. En contraste, el procedimiento que se llevó a cabo en la tesis utiliza información que contempla la escala de las deleciones al agruparlas por medio del sobrelape descrito en la metodología.

Comparar con un solo método es bastante incompleto, por lo que lo ideal sería extender la comparación con programas como `DELLY` (52) y `LUMPY` (36), los cuales se ajustan más a los tipos de datos que queremos analizar. Es decir, secuneciaciones WGS sin un tejido de comparación. De igual manera, sería conveniente comparar las deleciones con las bases de datos de variaciones poblacionales identificadas actualmente como gnomAD, dbSNP o COSMIC.

Apéndice A

Apéndice

A.1. Sección técnica del desarrollo

En los siguientes comandos se utiliza la forma de asignar variables que existen en los sistemas UNIX, esto quiere decir que las variables están representadas con el formato `${Variable}`, por ejemplo, `Enter : ${name}` donde `name` toma el valor de un nombre e indica que la implementación tiene que ser escrita así: `Enter : Edgar` sin el signo inicial de pesos; la configuración de los programas implementados en este proyecto son configurables por medio de la línea de comando. Por lo tanto, se le pueden asignar valores por medio de variables de la misma forma en que se utilizan en el texto. El inicio de la línea de comando siempre empieza con el carácter `$`, que indica la entrada de la terminal; la única sección que no sigue esta convención es la "Búsqueda en SRA", que sigue la convención de la búsqueda de la interfaz web.

A.1.1. Búsqueda en SRA

La interfaz web tiene la opción de hacer búsquedas avanzadas por medio de filtros. En este caso en particular se aplicaron los siguientes términos de búsqueda:

```
Human [ORGN] AND strategy wgs [FILT] AND Public [ACS] AND
    genomic [FILT] AND "Variable" [MBS]
```

Se pueden identificar perfectamente los parámetros determinados anteriormente, separados por el operador `AND`. El término `Human [ORGN]` corresponde a la búsqueda en datos en la especie humana; strategy `strategy wgs[FILT]` se refiere a la estrategia experimental "WGS"; el término `genomic [FILT]` restringe las secuenciaciones a datos genómicos.

Es necesario buscar la restricción de datos genómicos porque existen análisis enfocados a ARN o exones. El ARN se retrotranscribe generando lecturas de ADN en ambos casos, sin embargo, contienen información que no representa a todo el genoma.

Por último `"Variable"[MBS]` es la representación de la cantidad de pares de bases presentes en los experimentos de secuenciación, donde `Variable` es un número mayor a 3200, con la intención de obtener por lo menos una cobertura sobre el genoma humano.

Puesto que la interfaz web de NCBI no acepta ingresar rangos de búsqueda, se optó por usar una API web llamada "Entrez" (48), que permite automatizar las búsquedas de información en NCBI. De manera más específica, se usó la paquetería llamada "rEntrez" disponible en R, la cual lleva a cabo las mismas tareas que Entrez, con la ventaja de que aporta a la forma de automatizarse por medio de scripts en R.

A.1.2. Descarga

La descarga fue realizada por medio del protocolo FTP. Por lo que se utilizaron los siguientes comandos:

```
$ wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/
       sra/${name:0:3}${name:0:6}/${name}/${name}.sra
```

Donde la variable name corresponde al nombre de acceso de SRA, por ejemplo, si se tiene el nombre de acceso "ERR318658", entonces se tiene descarga por medio de:

```
$ wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/
       sra/ERR/ERR318/ERR318658/ERR318658.sra
```

La descarga arroja un archivo de extensión ".sra" con el nombre del experimento. Por ejemplo, "ERR318658" descarga el archivo "ERR318658.sra". Este archivo corresponde a un comprimido específico de NCBI que contiene la información necesaria para obtener los archivos crudos o procesados que se contienen en la base de datos. En el caso de este estudio, se obtuvieron los datos crudos por medio de la instrucción:

```
$ fastq-dump -split-files \
       -N ${inicio} \
       -X ${final} \
       -gzip \
       -O ${DIR}/${name}.sra
```

Donde `inicio` corresponde a la variable que indica el número de lectura en el inicio del intervalo a descargar. De manera análoga, `final` corresponde a la variable que indica el número de lectura en el final. La variable `DIR` es la dirección del fichero donde los archivos son descargados. El argumento `-gzip` indica que los archivos a descargar se guardan en formato comprimido "gz".

Los intervalos son necesarios en los experimentos de coberturas altas, dado que los programas de extracción y alineamiento tienen problemas con archivos grandes. Así que en dichos casos la extracción se lleva a cabo por intervalos de lecturas. El archivo resultante que contiene las lecturas es texto en codificación ASCII en formato FASTQ.

A.1.3. Calidad de experimentos

Para remover las subsecuencias de baja calidad de las lecturas se utilizó la herramienta `fastq_quality_trimmer` de "fastqx-Toolkit". Su uso es mostrado a continuación.

```
$ fastq_quality_trimmer -t ${quality} \
        -z -v \
        -l ${bases} \
        -o ${out_name}
```

Donde `out_name` es la variable que indica la dirección y nombre del archivo de salida. El formato de salida estará comprimido, a consecuencia del argumento `-v`. La variable `quality` corresponde a la calidad mínima requerida para las lecturas. De manera similar, `bases` es la longitud mínima de una lectura.

En la implementación que se hizo, el programa descartó bases con calidades Phred menores a 41, así como lecturas que contuvieran menos de 20 pares de bases. Los 20 pares de bases permitieron tener una alta probabilidad de un alineamiento sin ambigüedad (53).

A.1.4. Mapeo de lecturas

Se utilizó el programa `Segemehl` (24) para llevar a cabo el mapeo. Previamente al uso del programa, fue necesario construir un índice del genoma de referencia a utilizar, llamado "human.idx", el cual es necesario para mapear las lecturas al genoma humano de referencia. El uso del programa para esta tarea es el siguiente.

```
$ segemehl.x -x ${index}.idx \
        -d ${Genoma_Referencia} \
        --threads ${hilos}
```

El argumento `-x` indica la creación del índice con nombre `index`., con el cual es recomendable mantener la extensión "idx" en el nombre.

El segundo argumento corresponde al nombre y la ruta del archivo que contiene el genoma de referencia. En este caso, el genoma de referencia completo con nombre: "Homo sapiens.GRCh38.dna.primary assembly.fa"

Por último, se indica la cantidad de hilos a usar para la ejecución del programa con la variable `hilos`. De igual manera, dicho argumento se utilizará en el alineamiento.

Los hilos son la cantidad de tareas que se pueden realizar en distintos procesadores. De esta manera el tiempo de ejecución disminuye. El índice construido se puede utilizar para cualquier alineamiento posterior. El uso del programa para el mapeo de lecturas es el siguiente.

```
$ segemehl.x -i ${IdxSegemehl} \
        -S \
        -d ${Genoma_Referencia} \
        -q ${fastq} \
        -t ${hilos} \
        | samtools view -Sbh > ${alineamiento}.bam
```

La mayoría de los parámetros son similares al comando visto anteriormente. Los primeros dos argumentos corresponden al nombre, con ruta del archivo, del índice y genoma de referencia respectivamente. El tercer argumento `fastq` corresponde al archivo de extensión "fastq", el cual contiene las lecturas. Por último, es indicada la cantidad de hilos, como se explicó en el comando anterior. La salida del programa es un archivo en formato SAM. Para disminuir el uso de memoria, se guardó en formato BAM, el cual es un archivo binario que contiene la misma información. Este cambio de archivo se logra a través de la paquetería de `samtools` (38). Los argumentos `-Sbh` indican el tipo de conversión, así como la inclusión de las cabeceras del SAM.

A.1.5. Detección de gaps

El programa recibe las líneas del SAM por `stdin`. Como salida, escribe cada una de las deleciones reportadas en archivos con formato BED.

Como requerimiento de los próximos análisis, es necesario crear un archivo por cada uno de los cromosomas reportados en el SAM. Para ello, en cada línea se revisa el cromosoma al cual corresponde a la línea del SAM. Con ello, las deleciones se guardan en el archivo llamado a partir de un nombre base y el nombre del cromosoma.

El formato del nombre de los archivos de salida es `${base_name}_${cromosoma}.bed`. Donde `base_name` es ingresado por el usuario.

```
$ SAM_parce_GAP ${name} ${Flanco}
```

La variable `name` es el nombre base con el cual se nombran los archivos de salida. La variable `Flanco` indica la cantidad de pares de bases mínima que tiene que tener cada flanco de un gap para ser reportado.

A.1.6. Filtros

A.1.6.1. Regiones repetidas en RepeatMasker

Este programa lee el BED de entrada por por `stdin` y despliega un BED filtrado por `stdout`. La ejecución del programa se lleva a cabo de la siguiente manera:

```
$ RMFilter ${Chromosome}.bin ${lines} ${qbases} ${percentage}
```

`Chromosome` es la dirección y nombre del archivo binario que contiene las coordenadas de RepeatMasker para un cromosoma determinado. Este archivo fue creado previamente a partir de la base descargada de la página oficial de RepeatMasker. A partir del archivo descargado se extrajeron las coordenadas en distintos archivos por cada cromosoma, y se guardaron en formato binario.

El segundo parámetro `lines` es el número que representa la cantidad de coordenadas en la base de datos, dicho dato fue generado a la par de los binarios. El parámetro `qbases` corresponde a la cantidad de bases en cada extremo, que definen las bases aledañas. Por último, el parámetro de `percentage` es el que indica el umbral máximo para conservar una deleción.

A.1.6.2. Recurrencia de deleciones en un mismo experimento

Para poder usar el programa se utiliza el comando de la siguiente manera:

```
$ Del_Overlap ${name}.bed \
        ${score_min} \
        ${overlap_percentaje}
```

Donde `name` es la variable del nombre del archivo BED de un solo cromosoma; este archivo tiene que estar ordenado como condición para el funcionamiento del programa. El valor de `score_min` es la variable que determina el mínimo de recurrencias necesarias para reportar la deleción. La variable `overlap_percentaje` es el valor en cuanto a porcentaje mínimo de cobertura para determinar un sobrelape. La salida se obtiene por `stdout` en formato BED.

A.1.6.3. Filtro somático

Para este análisis se requiere utilizar tres aplicaciones distintas, tal y como se expone a continuación:

genomecov

Primero se debe obtener la cobertura por base en el genoma, por lo que se utilizó `genomecov`, con los parámetros `-split` y `-bg`.. El primer argumento permite considerar las regiones que no aportan a la cobertura dentro del SAM. El segundo parámetro permite una representación en formato BED de las regiones con la misma cobertura, con el fin de usar menor cantidad de memoria.

```
$ bedtools genomecov -split -bg -ibam ${name}.bam
```

Donde la variable `name` es claramente el nombre del archivo. A su vez, requirió de un programa intermedio para llevar a cabo la cuenta de las regiones de interés y conformar la base de datos de las mismas.

Coverage_count

Después se requiere generar la base de datos de la cobertura por regiones de deleción y flancos. Para lo cual este programa recibe la salida de `genomecov` y el BED de las regiones de interés. En necesario cuidar que ambas posean el mismo tipo de ordenamiento.

La entrada de `Coverage_count` es por `stdin`, y la salida por `stdout`.. Así que su implementación es la siguiente, donde `name` es el nombre del BED con las regiones de interés.

```
$ Coverage_count ${name}.bed \
```

Allelic_Filter

Por último se aplica el filtro, para lo cual se desarrolló `Allelic_Filter`, que recibe el archivo de las deleciones detectadas, la salida de `Coverage_count` y el umbral de las variaciones de cobertura. A su vez regresa a las deleciones filtradas con su respectiva fracción de cobertura.

```
$ Allelic_Filter ${name}_Overlap.bed ${name}_DB.bed 0.025
```

A.1.6.4. Recurrencia de deleciones entre individuos

Para este filtro se utilizó una de las herramientas de bedtools `bedtools` llamada `Intersect` (51).

```
$ bedtools intersect -wa -wb \
                -a ${Base} \
                -b ${Compared} \
                -sorted \
                -filenames \
                -f ${fraction} -r
```

Los argumentos `-wa -wb` permiten mantener las coordenadas de los registros originales en cada deleción reportada. La variable `Base` es el archivo en formato BED que se toma como referencia en la comparación. Al mismo tiempo, la variable `Compared` es la variable que contiene los archivos contra los cuales se realiza la comparación. El argumento `-sorted` indica que los archivos de entrada están ordenados. Cabe anotar que debido a los pasos anteriores, los archivos ya están ordenados. Para finalizar, la variable `fraction` indica la fracción de cobertura mínima que se requiere que contenga el sobrelape para ser reportado. Dicho criterio se aplica de manera recíproca, con el argumento `-r`.

A.2. Representación de las deleciones en el genoma

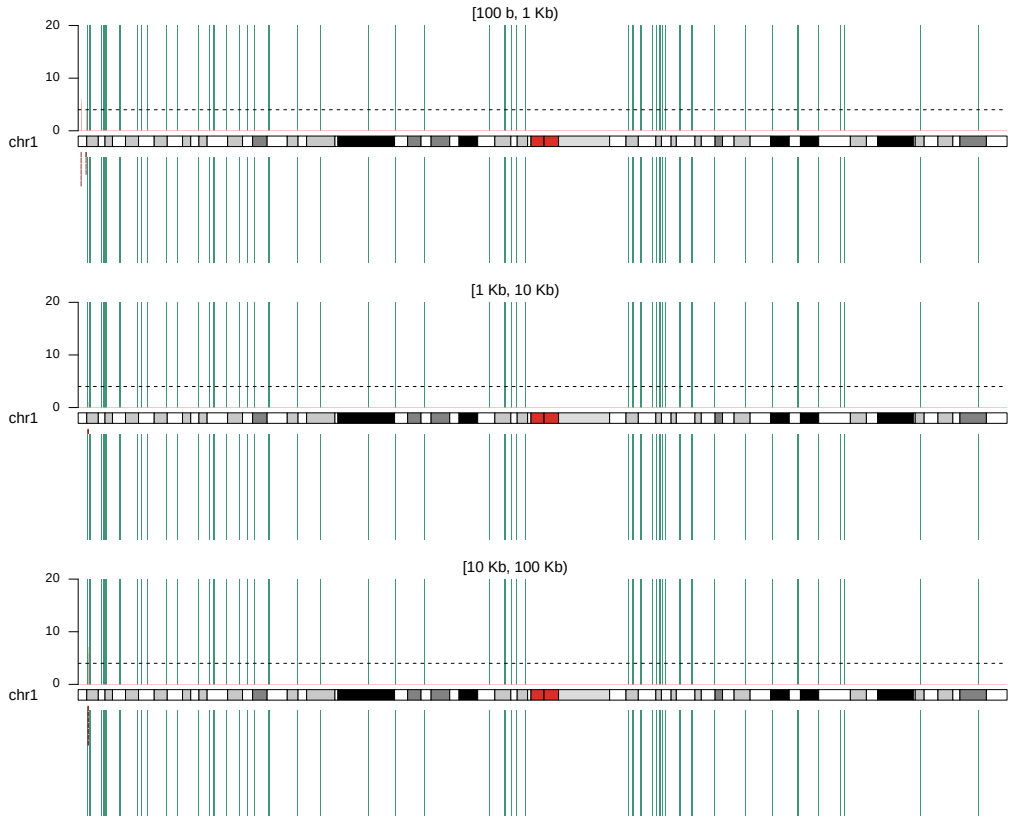
Para tener una idea de la forma en que se distribuyen las deleciones recurrentes en los cromosomas, estas fueron agrupadas en los siguientes intervalos de longitudes.

- $[1b, 10b)$
- $[10b, 100b)$
- $[100b, 1Kb)$
- $[1Kb, 10Kb)$
- $[10Kb, 100Kb)$
- $[100Kb, 1Mb)$
- $[1Mb, 10Mb)$
- $[10Mb, 100Mb)$
- $[100Mb, 500Mb)$

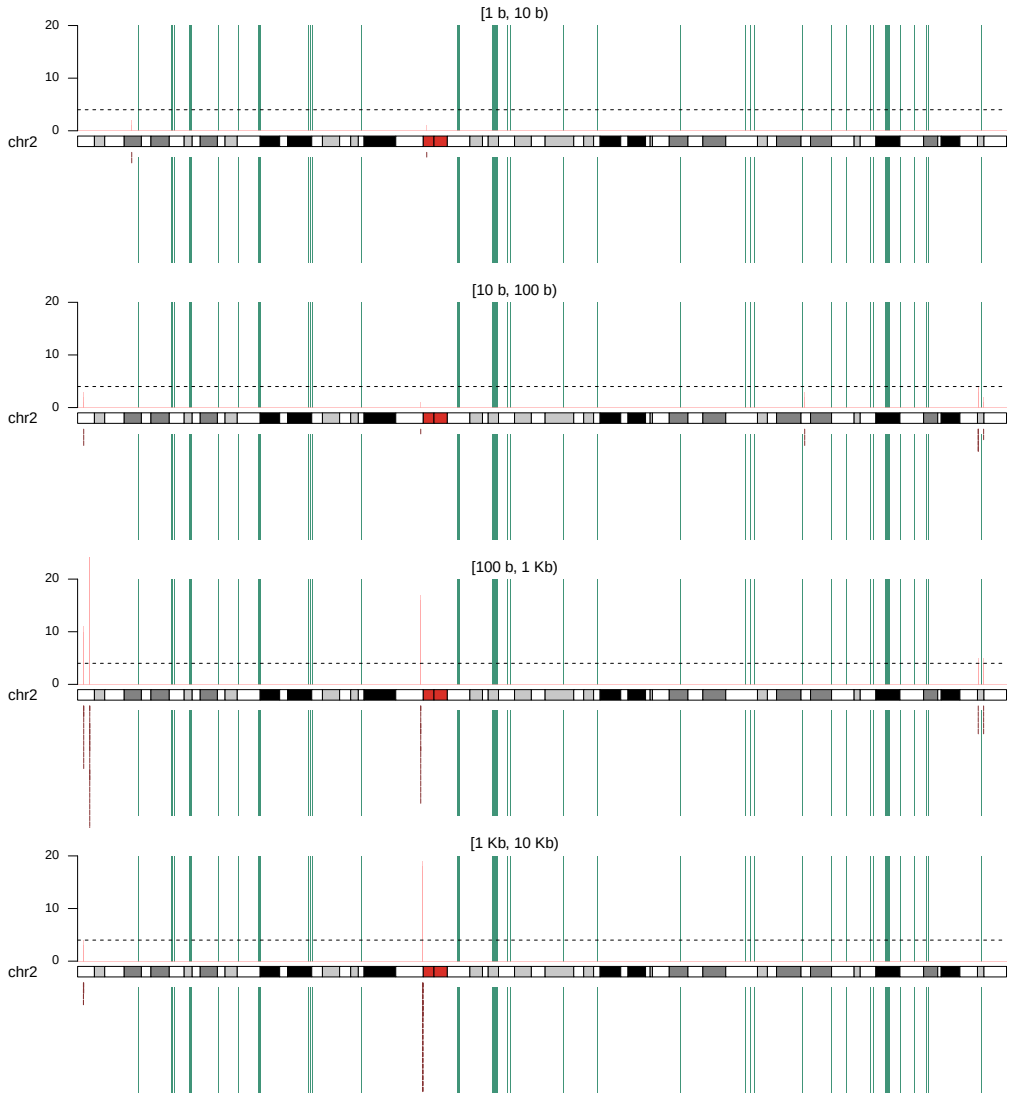
Fueron realizados gráficos por medio de `KaryoploteR` (18) por cada subconjunto, y a su vez por cada cromosoma. Cada gráfico se compone de un ideograma como referencia, el cual representa el cariotipo del cromosoma correspondiente, en la parte inferior son representadas las deleciones individuales, y en la parte superior se presenta un histograma que refleja la cantidad de deleciones que intersectan a lo largo del cromosoma. Hay que tener en cuenta que las escalas pueden variar, sin embargo, existe una línea de referencia en el histograma al nivel de cuatro. Esto se debe a que una deleción se cataloga como recurrente en una región si la frecuencia es mayor o igual a cuatro. No obstante, una mayor frecuencia indica mayor variabilidad, y por lo tanto una mayor dificultad de comprobación.

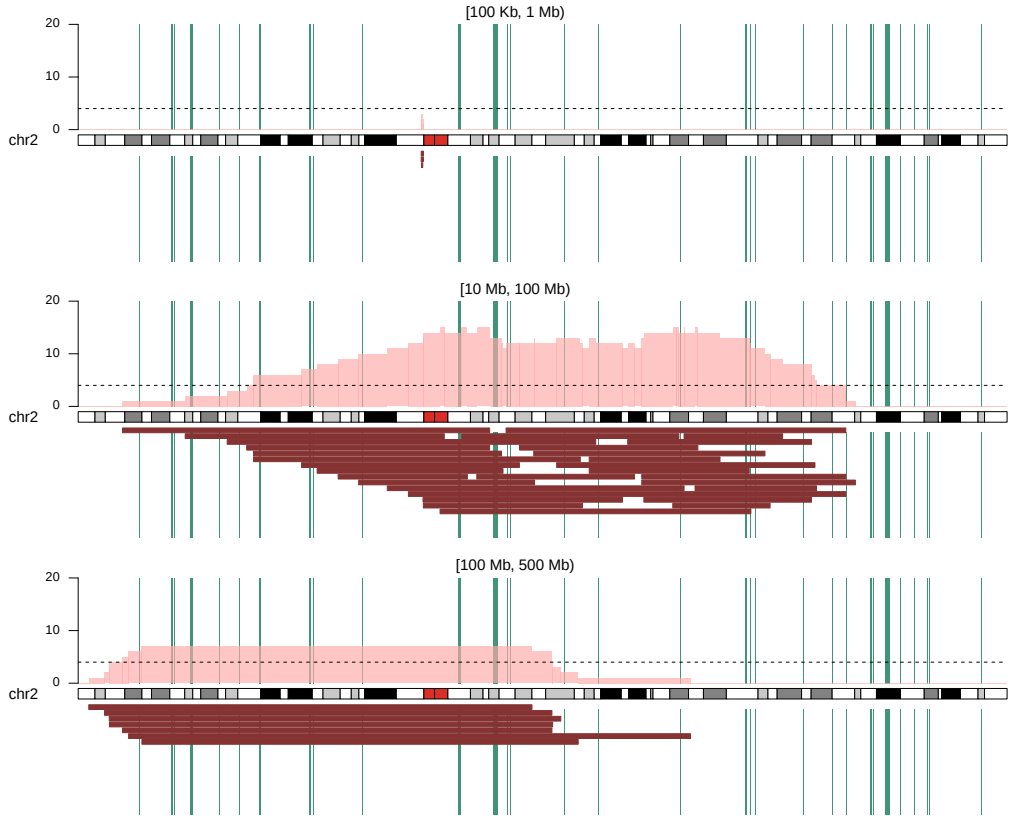
En cada uno de los gráficos se tienen las referencias de las regiones que corresponden a los genes relacionados a cancer, las cuales son representadas como sombras en un tono verde. A continuación se muestran los gráficos separados en cromosomas. Los intervalos faltantes se debe a que no existían deleciones compartidas entre experimentos dentro de dicho intervalo.

A.2.1. Cromosoma 1

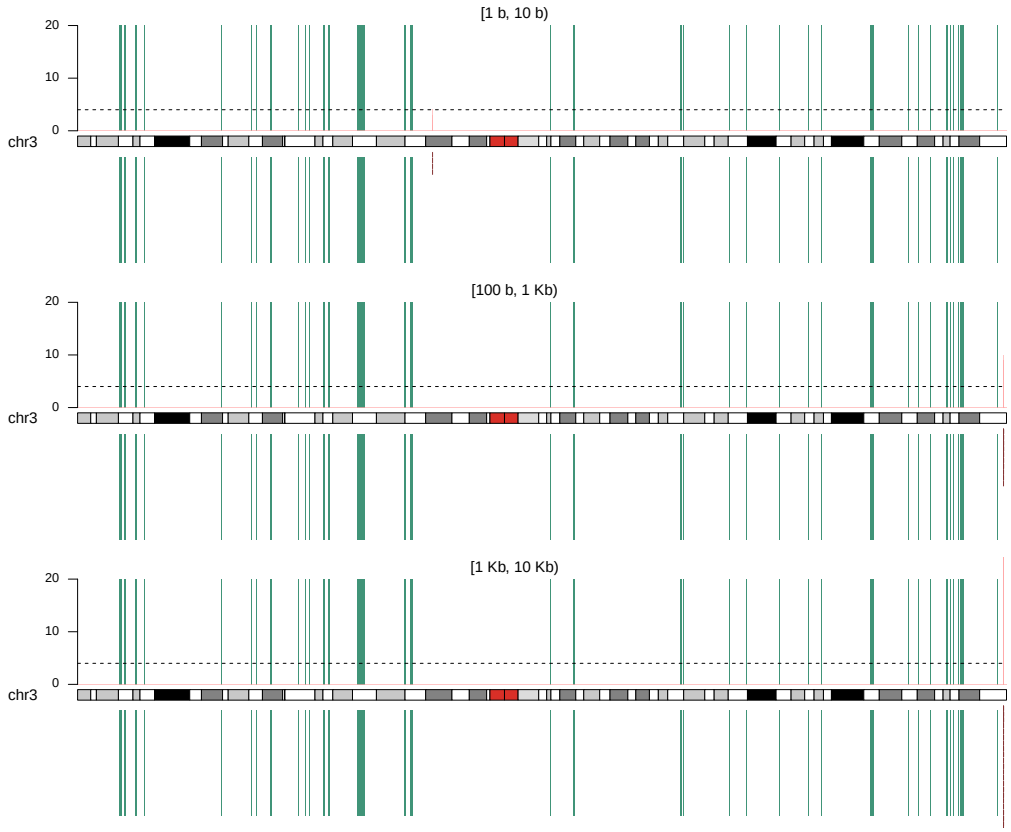


A.2.2. Cromosoma 2

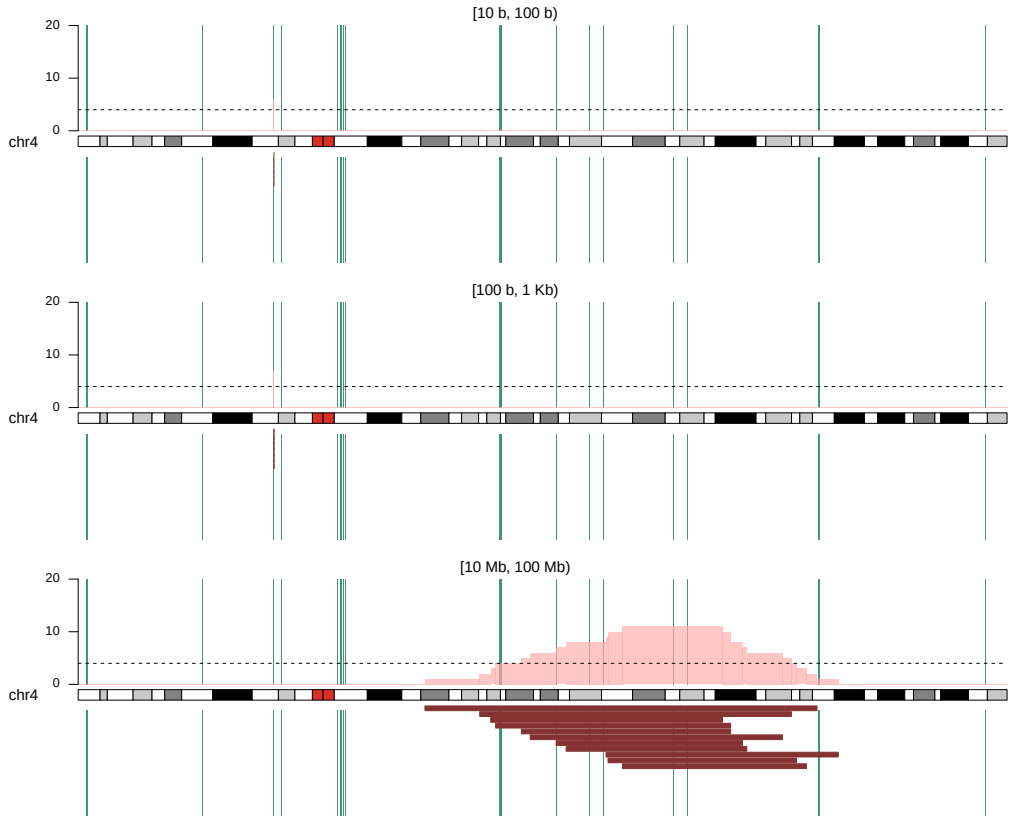



[100 Kb, 1 Mb)
20
10
0
chr2
[10 Mb, 100 Mb)
20
10
0
chr2
[100 Mb, 500 Mb)
20
10
0
chr2

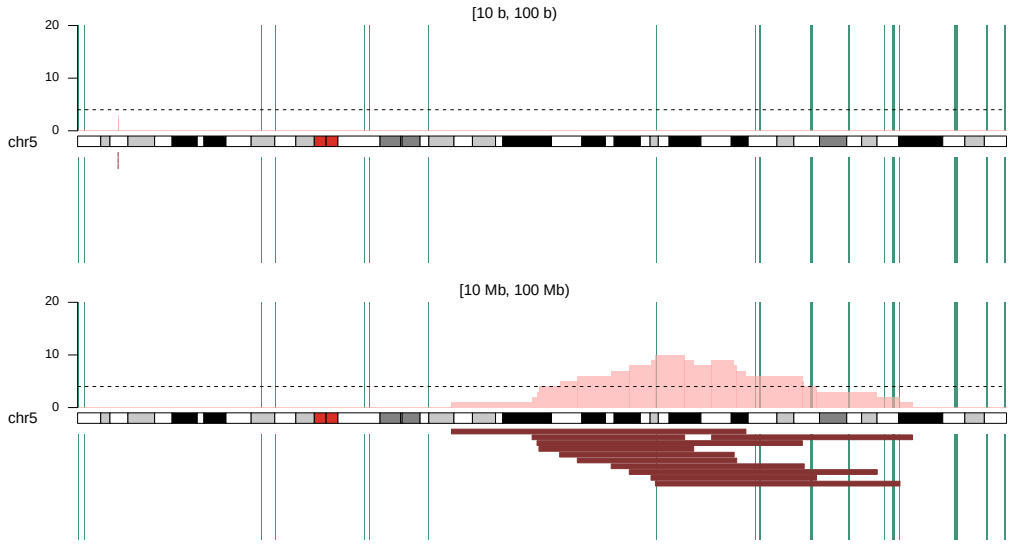
A.2.3. Cromosoma 3



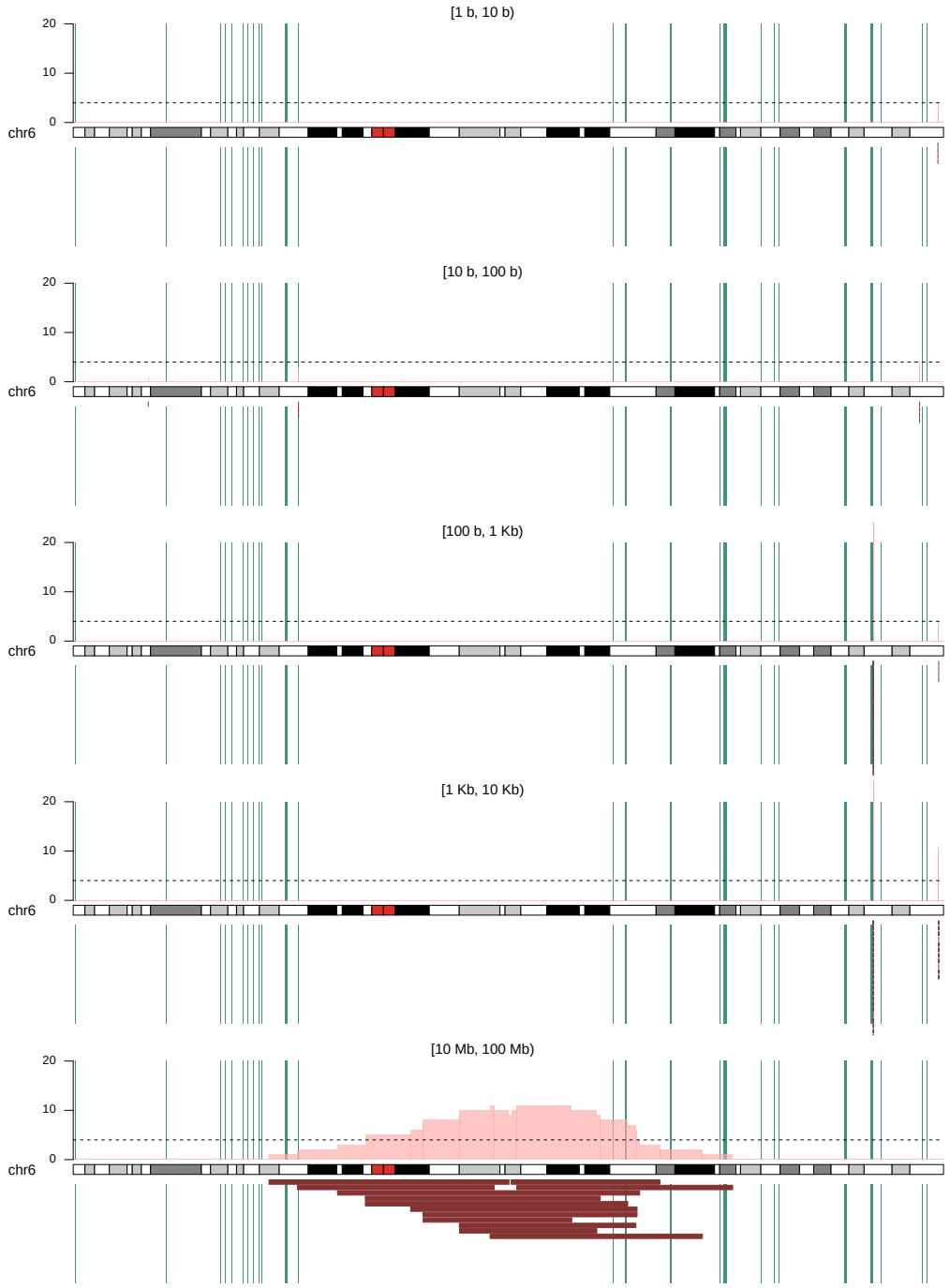
A.2.4. Cromosoma 4



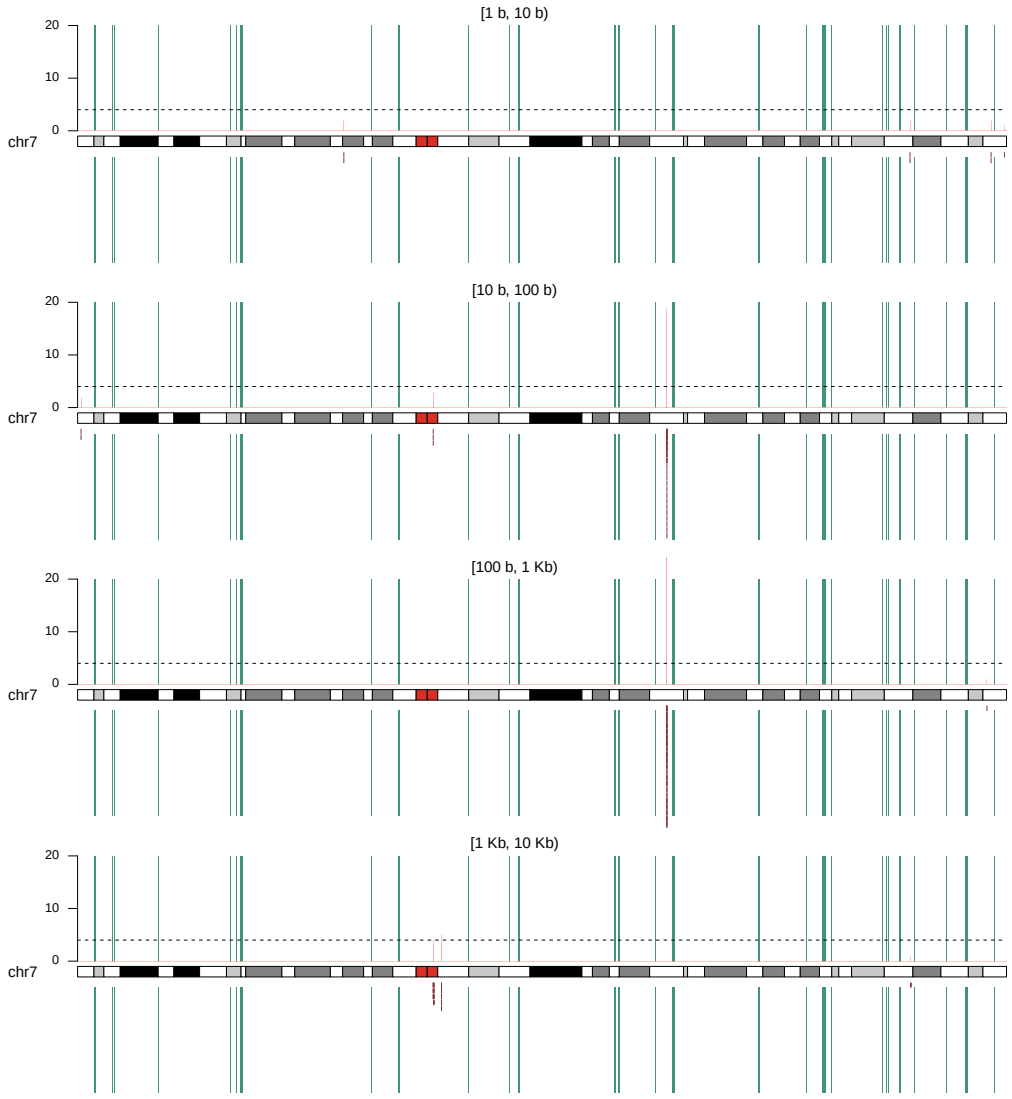
A.2.5. Cromosoma 5

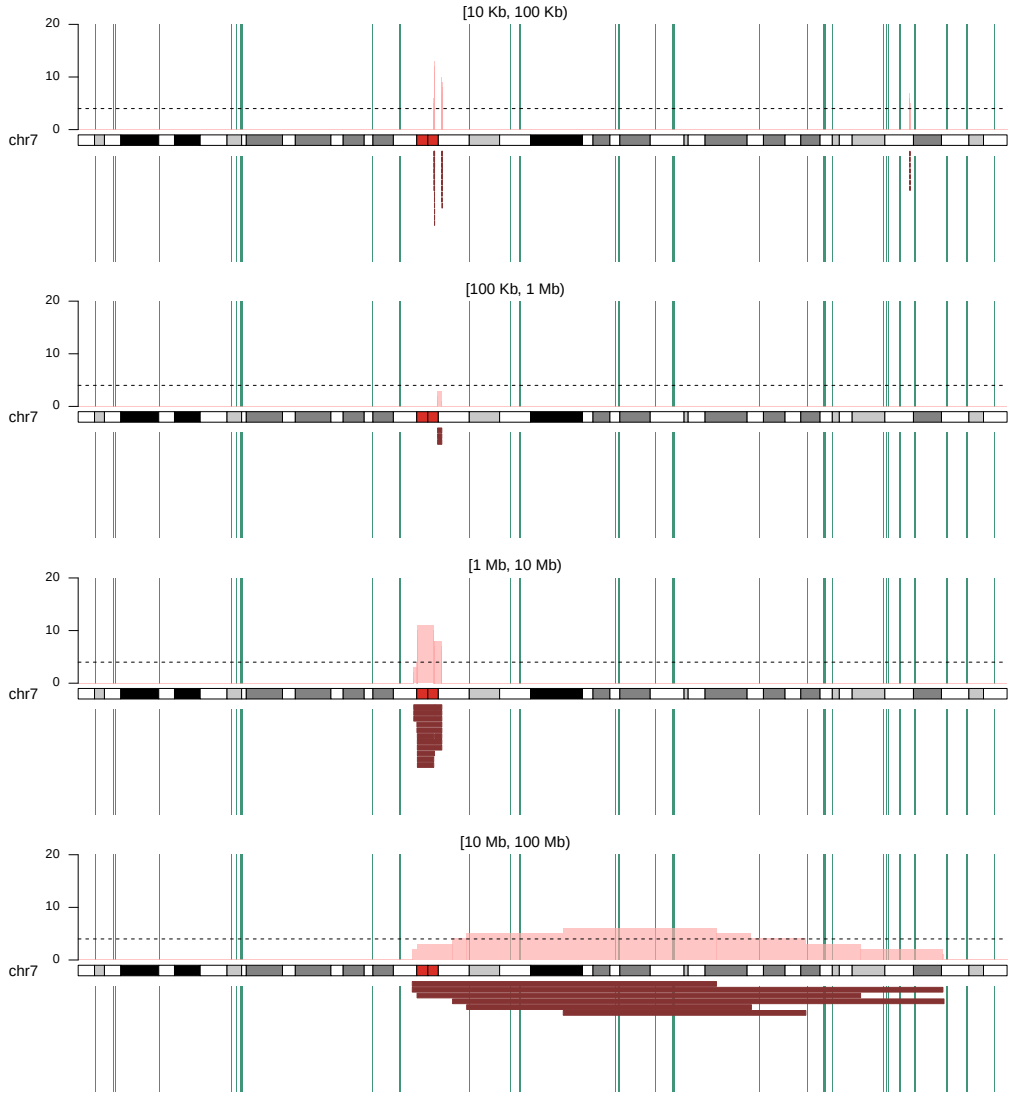


A.2.6. Cromosoma 6

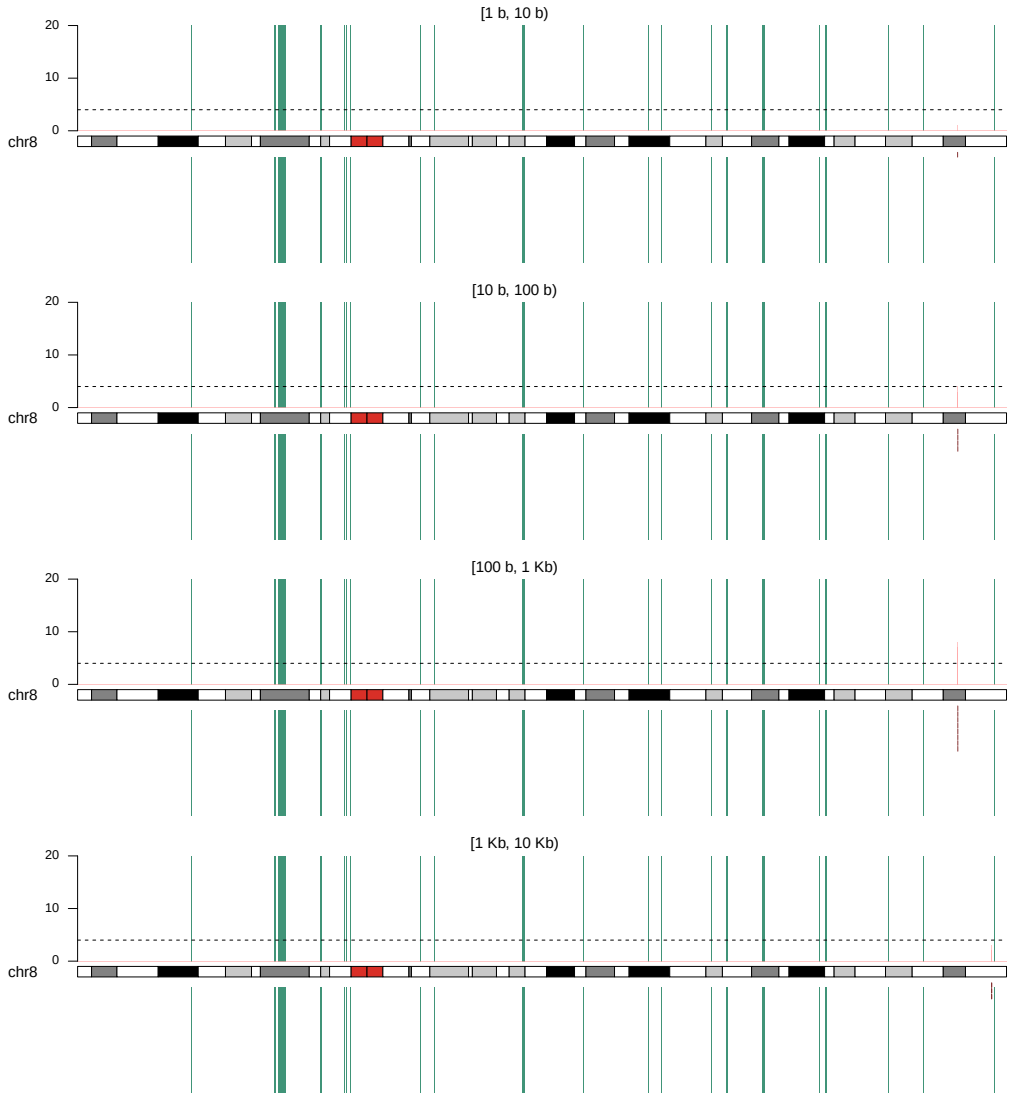


A.2.7. Cromosoma 7

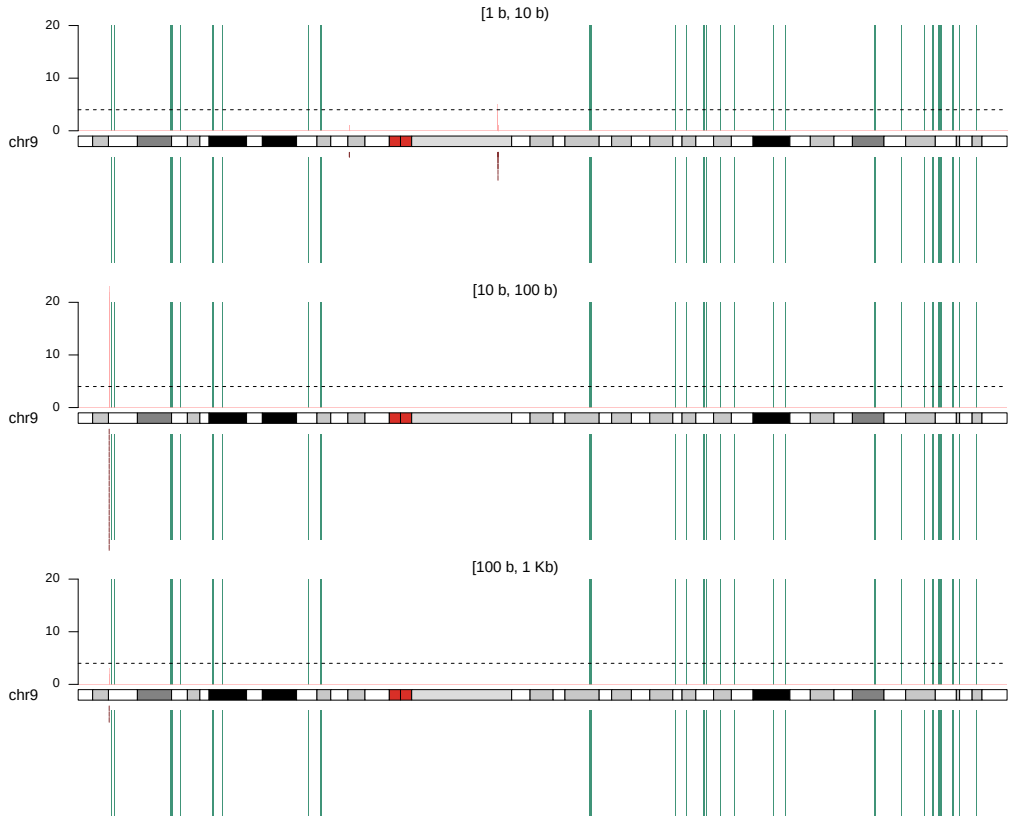


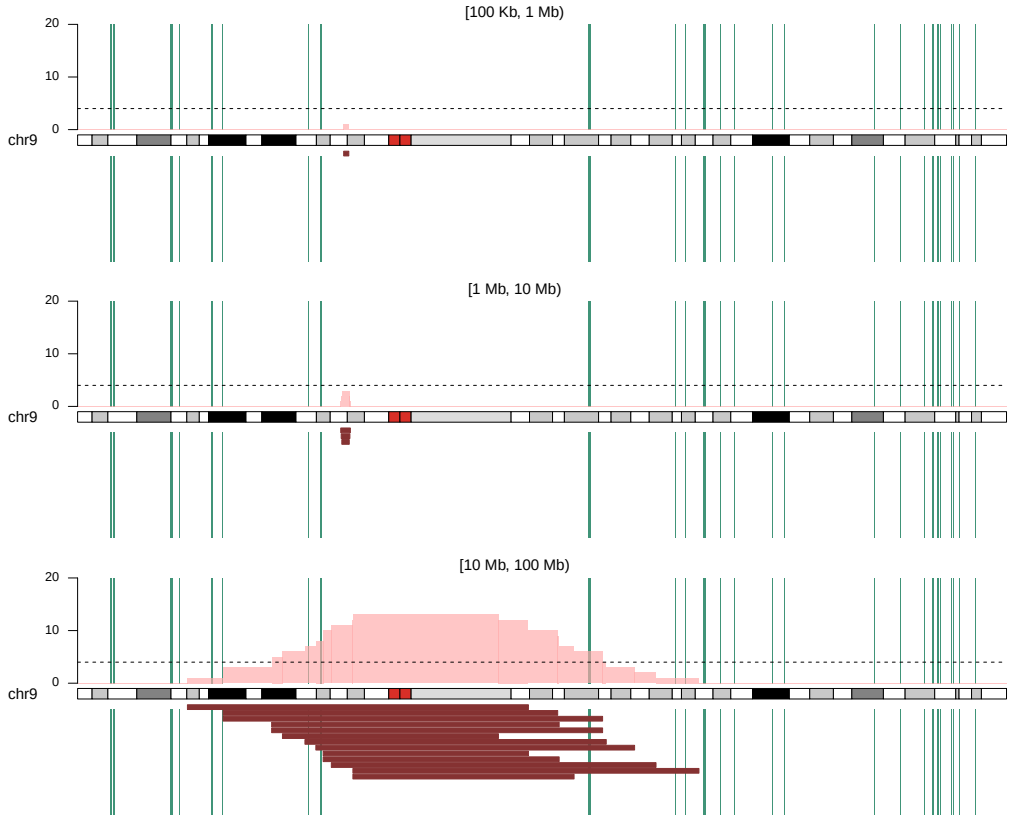

[10 Kb, 100 Kb)
20
10
0
chr7
[100 Kb, 1 Mb)
20
10
0
chr7
[1 Mb, 10 Mb)
20
10
0
chr7
[10 Mb, 100 Mb)
20
10
0
chr7

A.2.8. Cromosoma 8

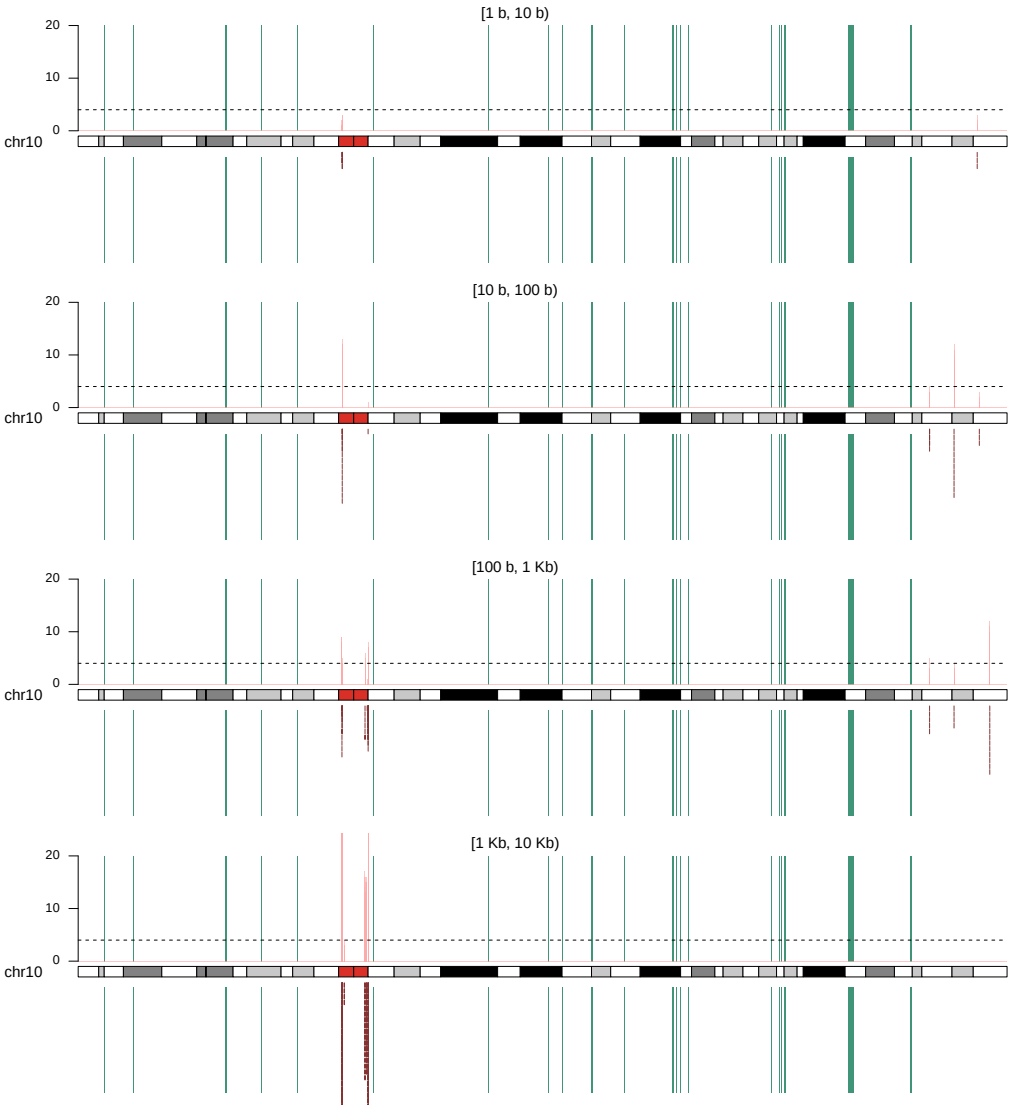


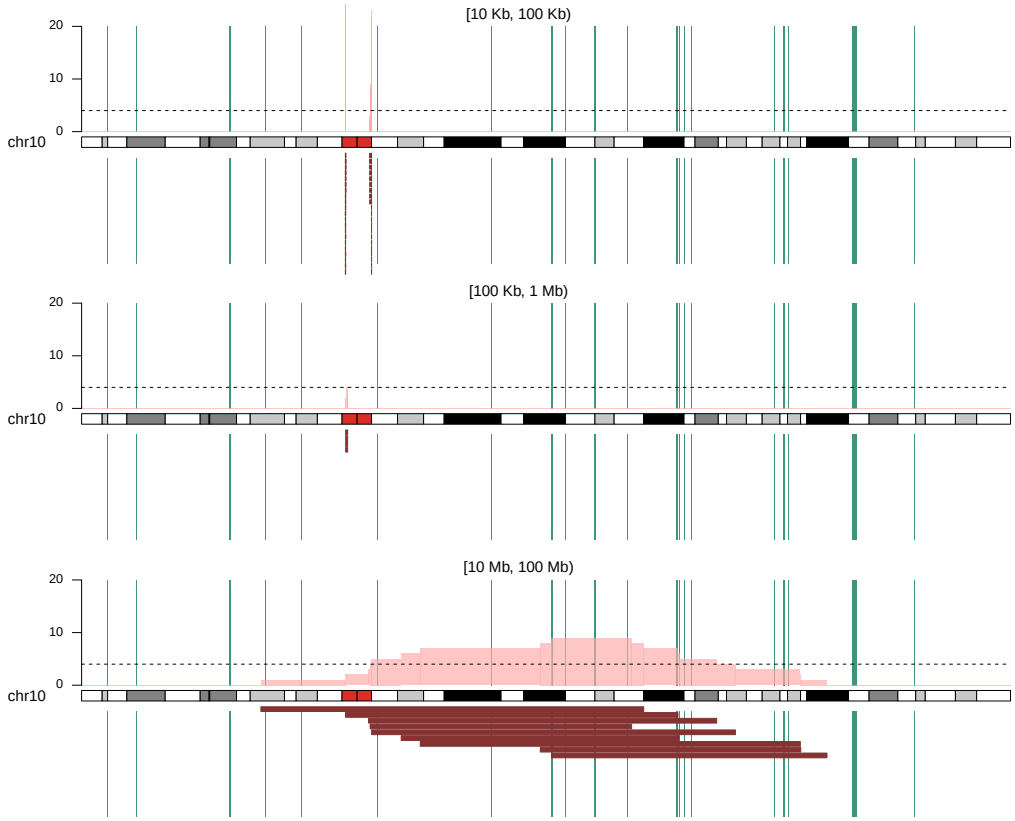
A.2.9. Cromosoma 9



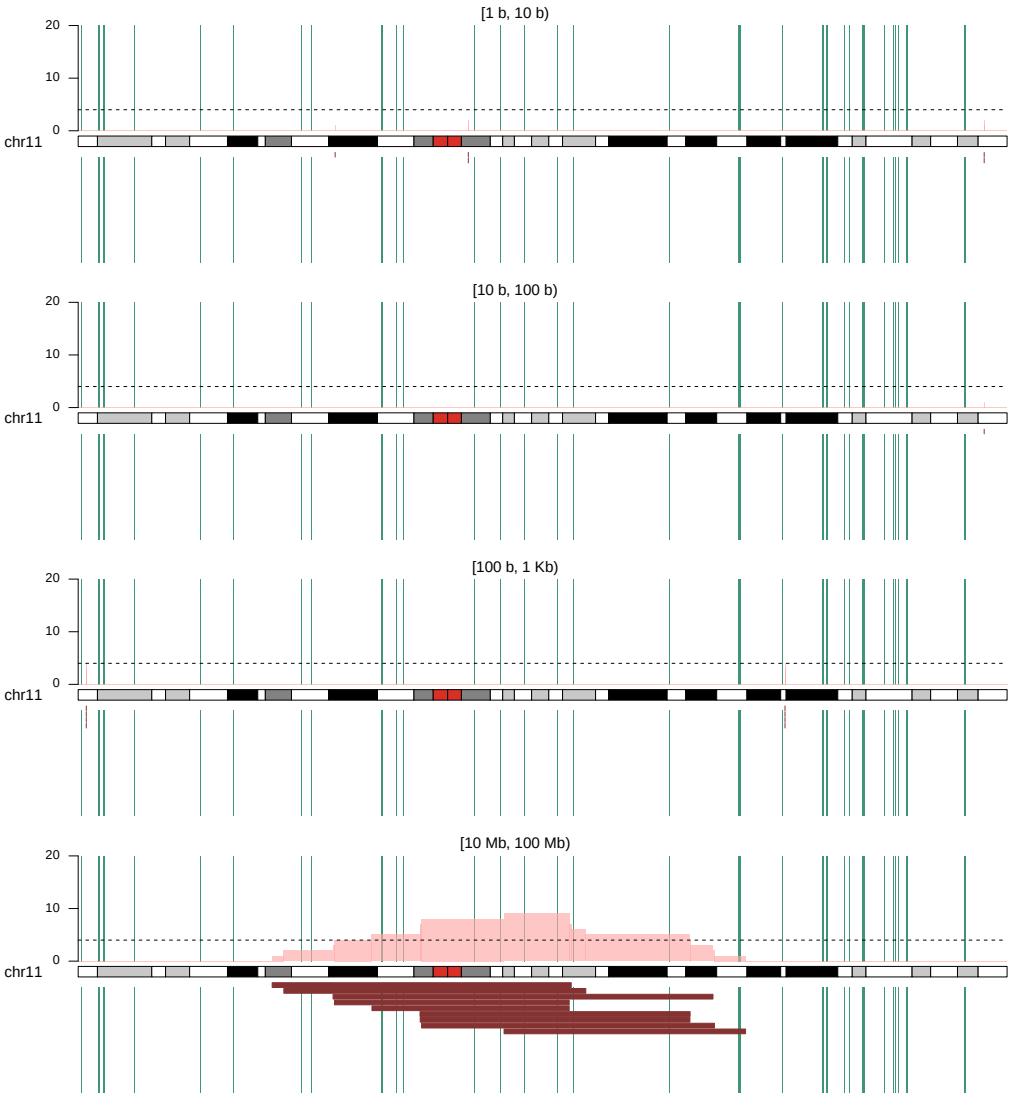

[100 Kb, 1 Mb)
20
10
0
chr9
[1 Mb, 10 Mb)
20
10
0
chr9
[10 Mb, 100 Mb)
20
10
0
chr9

A.2.10. Cromosoma 10

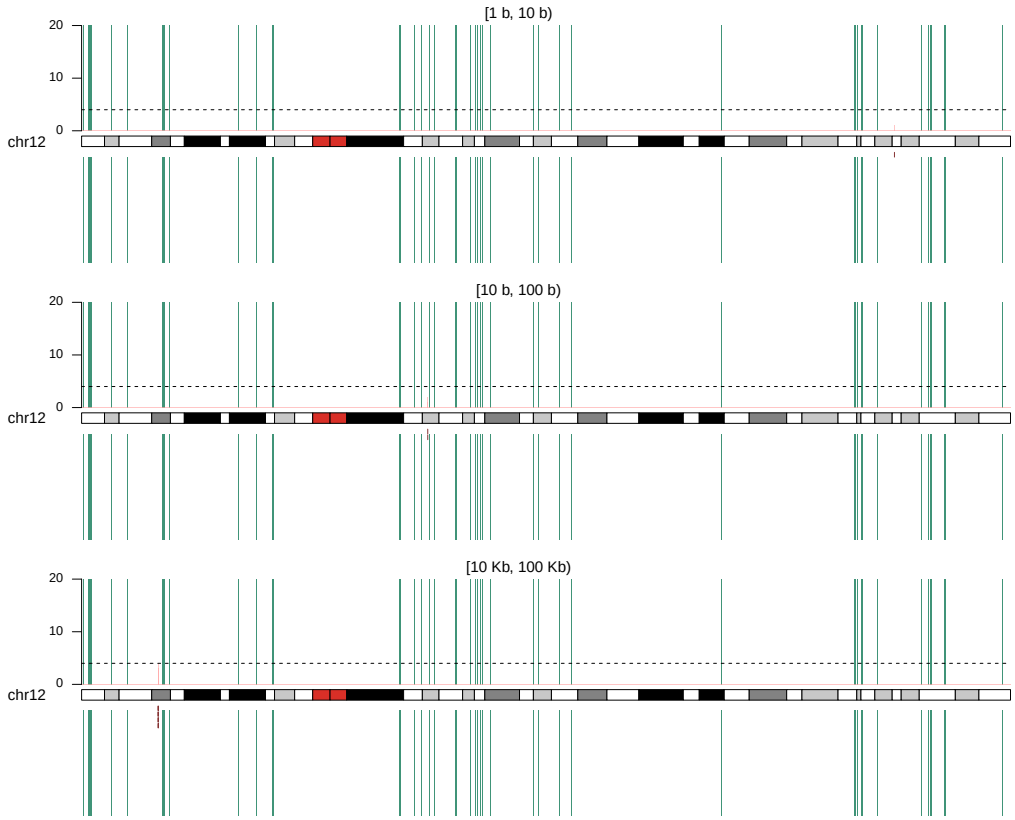



[10 Kb, 100 Kb)
20
10
0
chr10
[100 Kb, 1 Mb)
20
10
0
chr10
[10 Mb, 100 Mb)
20
10
0
chr10

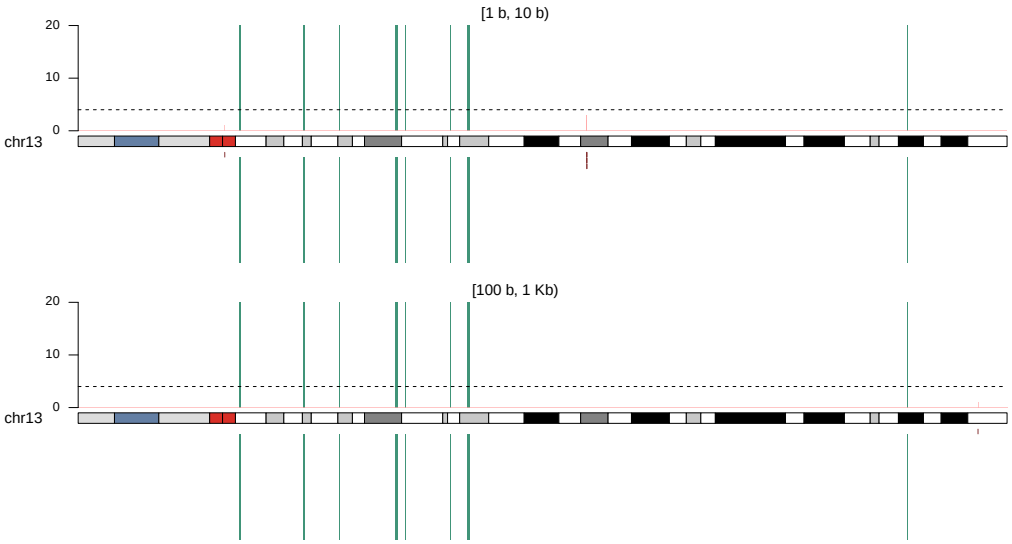
A.2.11. Cromosoma 11



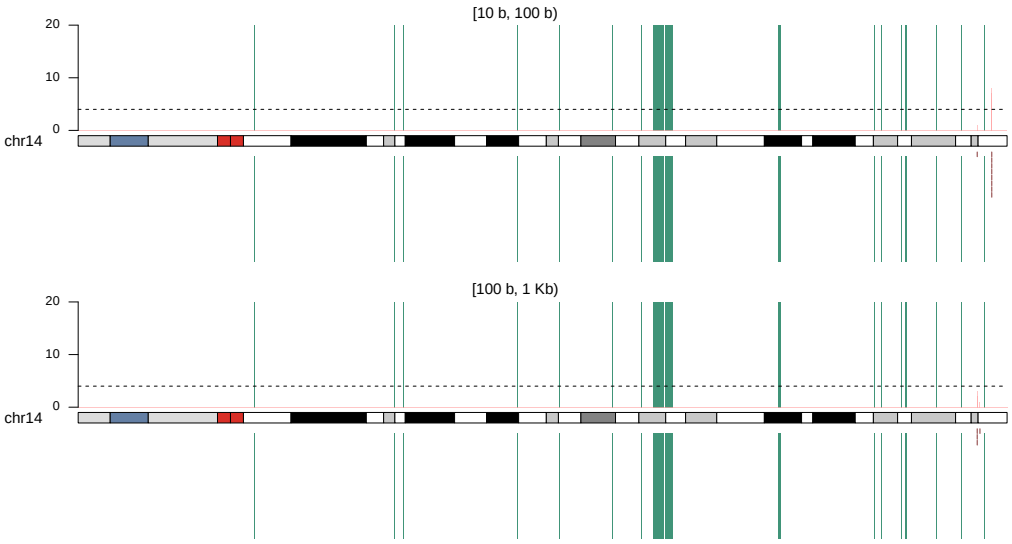
A.2.12. Cromosoma 12



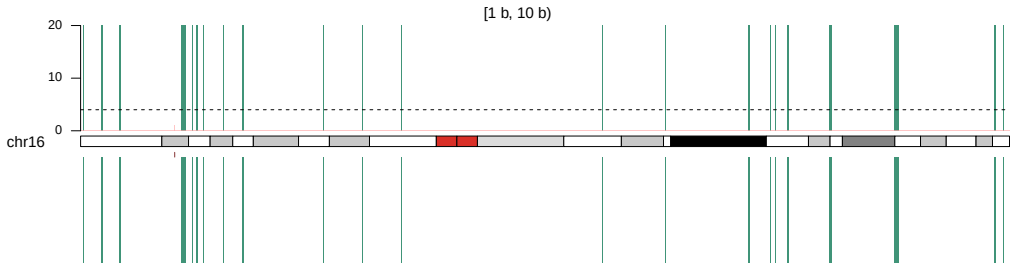
A.2.13. Cromosoma 13



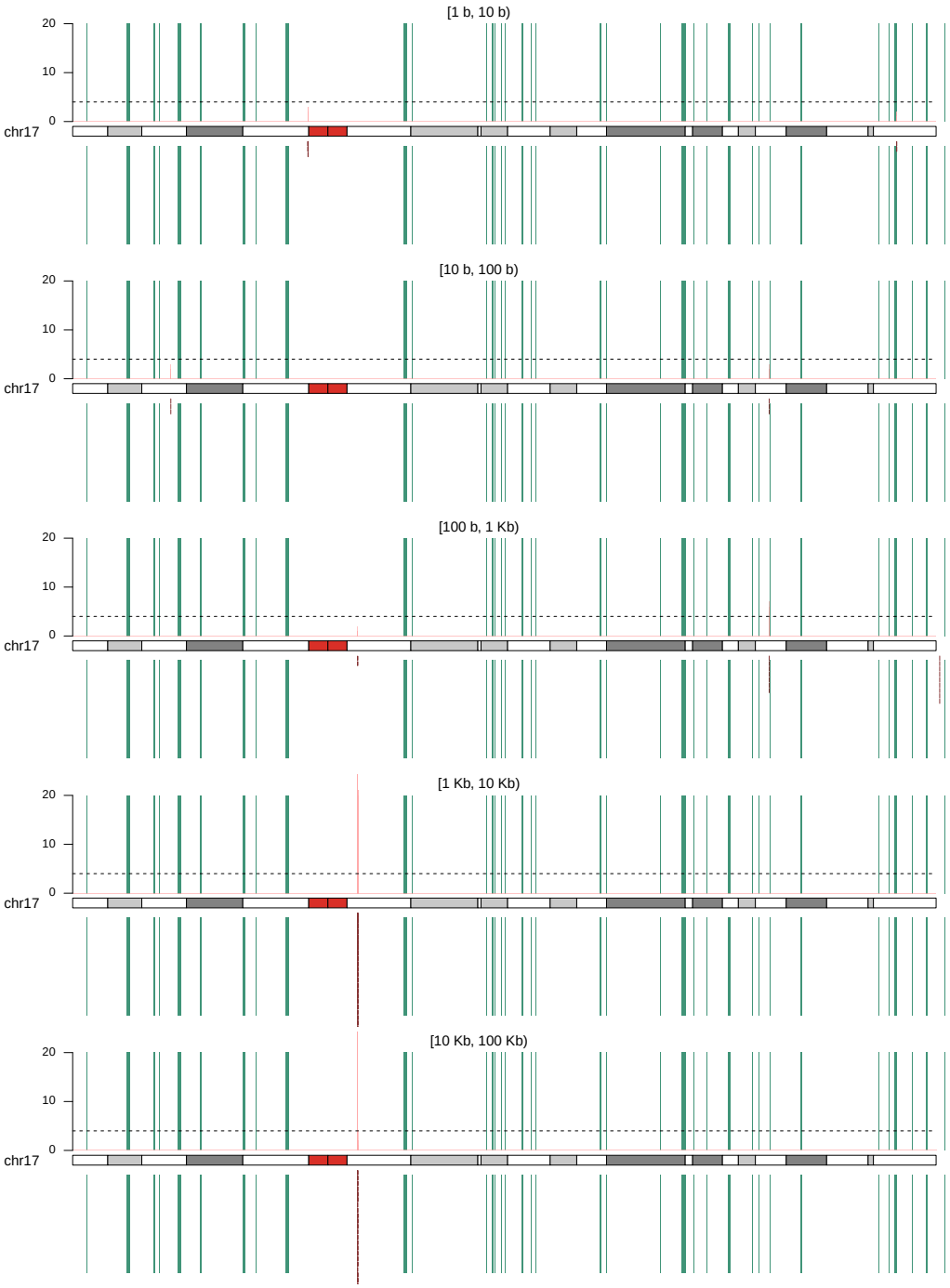
A.2.14. Cromosoma 14



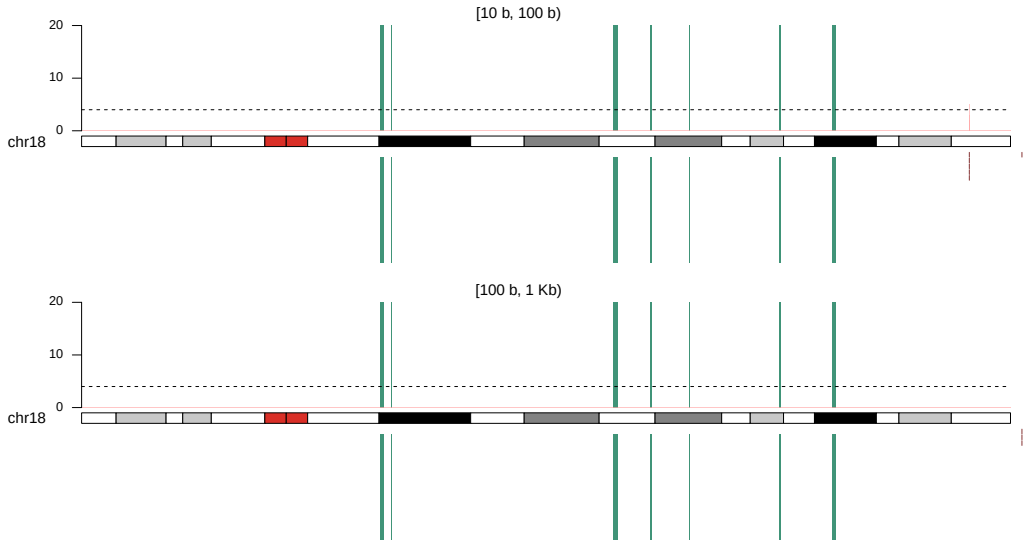
A.2.15. Cromosoma 16



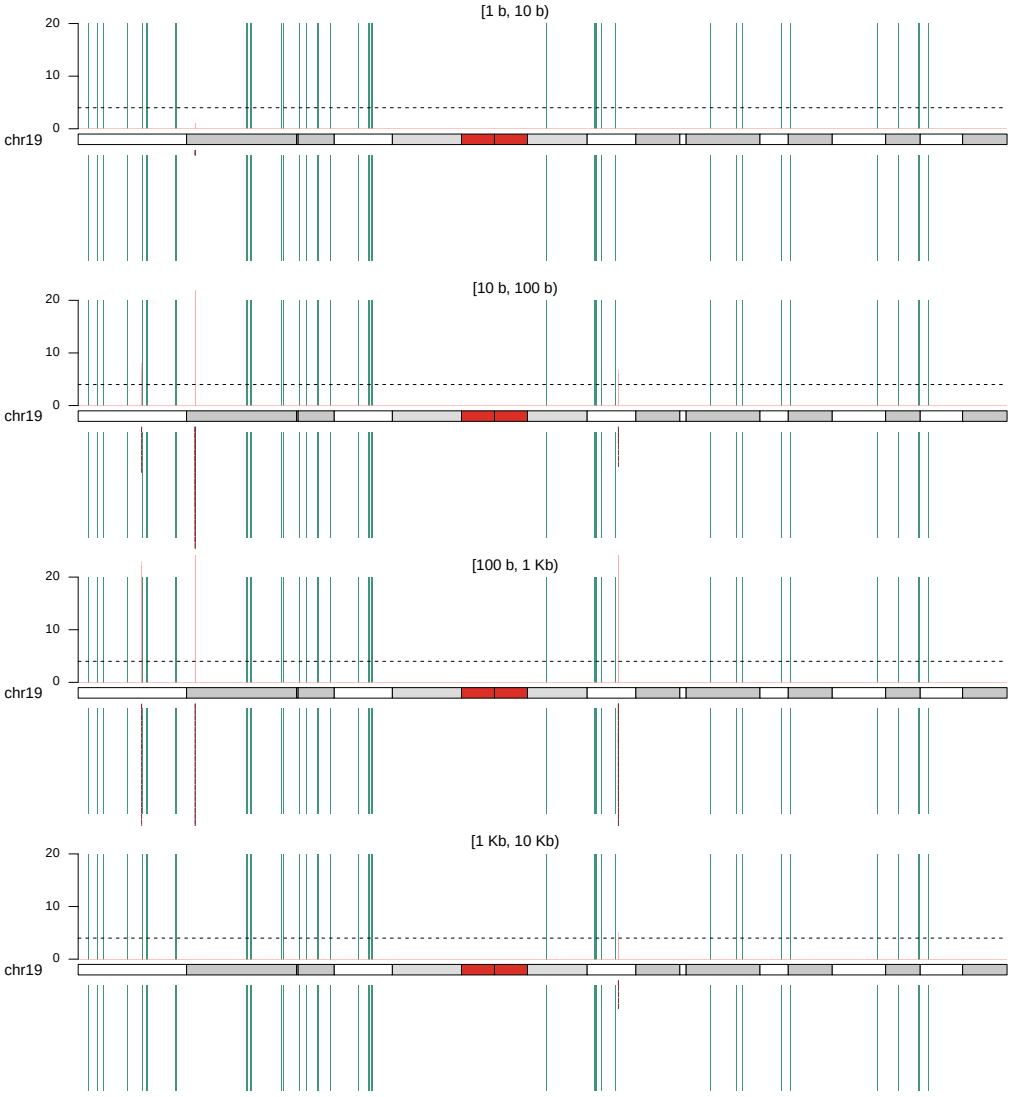
A.2.16. Cromosoma 17



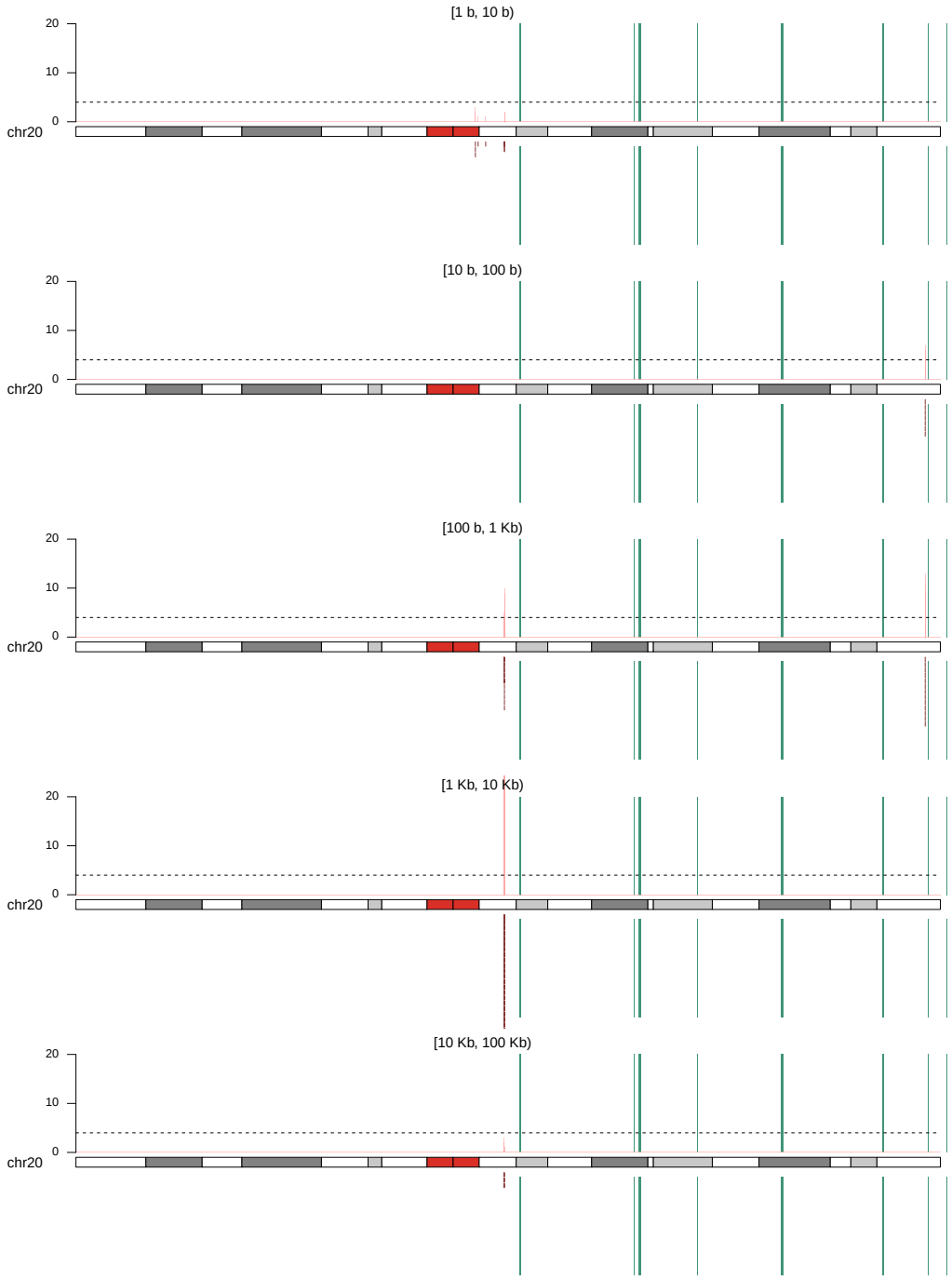
A.2.17. Cromosoma 18



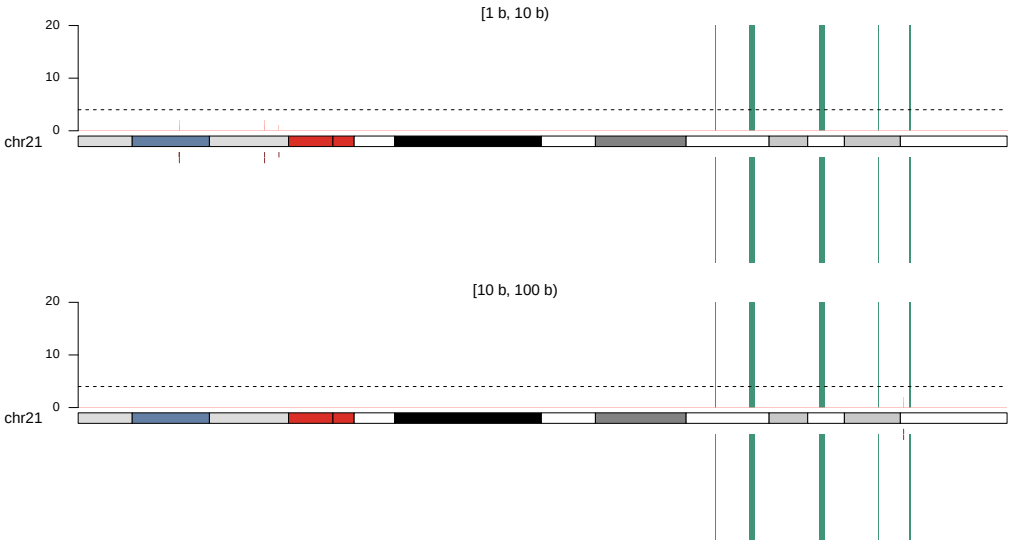
A.2.18. Cromosoma 19



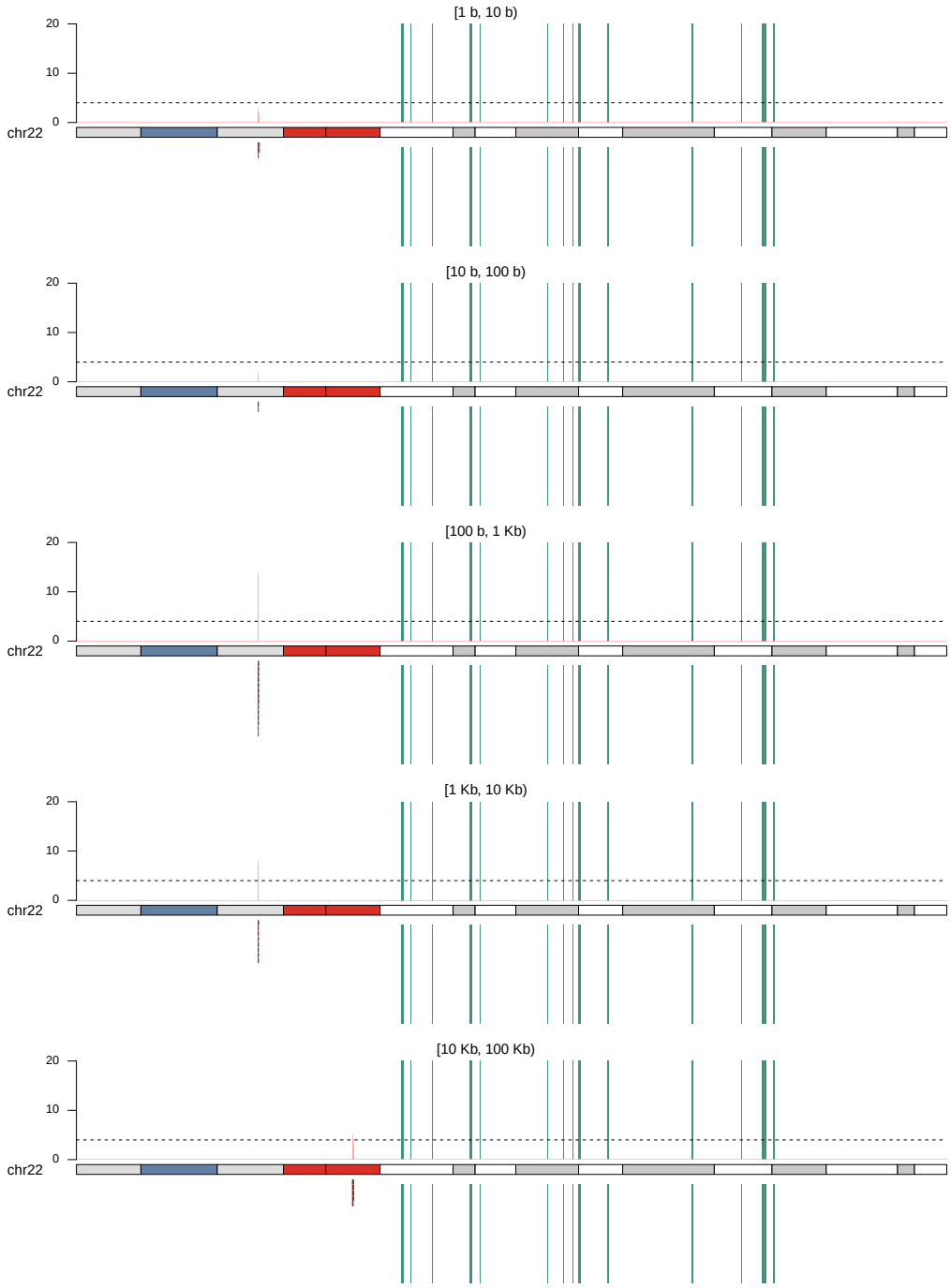
A.2.19. Cromosoma 20



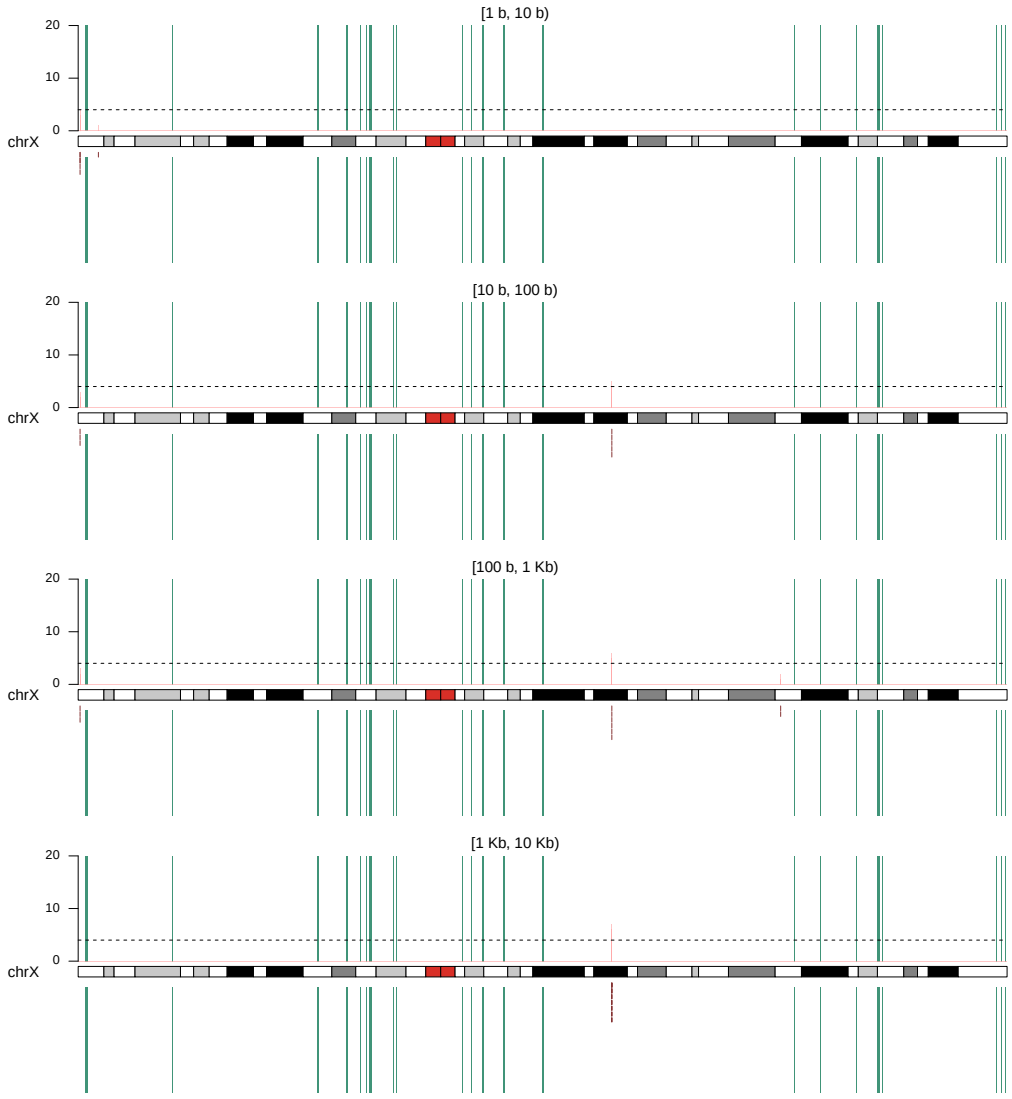
A.2.20. Cromosoma 21



A.2.21. Cromosoma 22



A.2.22. Cromosoma X



Bibliografía

[1] **Gene SEPT9 [Internet]**, 2004. Available from: `https://www.ncbi.nlm.nih.gov/gene/10801`.

[2] Joseph D Baugher, Benjamin D Baugher, Matthew D Shirley, and Jonathan Pevsner. **Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method**. *BMC Genomics*, **14**(1):1, 2013. Available from: `BMCGenomics`.

[3] Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. **Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data**. *Bioinformatics*, **28**(3):423–425, 2012.

[4] Coen Bron and Joep Kerbosch. **Algorithm 457: finding all cliques of an undirected graph**. *Communications of the ACM*, **16**(9):575–577, 1973.

[5] Stuart Brown. *Next-Generation DNA Sequencing Informatics, Second Edition.* 2015.

[6] A. Malcolm Campbell and Laurie J. Heyer. *Discovering genomics, proteomics, and bioinformatics.* CSHL Press Pearson/Benjamin Cummings, San Francisco, 2007.

[7] Ken Chen, John W Wallis, Michael D Mclellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D Mcgrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and R Elaine. **Breakdancer**. **6**(9):677–681, 2013.

[8] Mengjie Chen, Murat Gunel, and Hongyu Zhao. **SomatiCA: Identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data**. *PLoS ONE*, **8**(11), 2013.

[9] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Research*, **38**(6):1767–1771, 2009.

[10] Richard Coico. *Immunology : a short course*. John Wiley & Sons Inc, Chichester, West Sussex, UK Hoboken, NJ, 2015.

[11] The 1000 Genomes Project Consortium. **A global reference for human genetic variation**. *Nature*, **526**:68–74, 2015.

[12] Mircea Cretu Stancu, Markus J. Van Roosmalen, Ivo Renkens, Marleen M. Nieboer, Sjors Middelkamp, Joep De Ligt, Giulia Pregno, Daniela Giachino, Giorgia Mandrile, Jose Espejo Valle-Inclan, Jerome Korzelius, Ewart De Bruijn, Edwin Cuppen, Michael E. Talkowski, Tobias Marschall, Jeroen De Ridder, and Wigard P. Kloosterman. **Mapping and phasing of structural variation in patient genomes using nanopore sequencing**. *Nature Communications*, **8**(1):1–13, 2017. Available from: `http://dx.doi.org/10.1038/s41467-017-01343-4`.

[13] Sergio Ulhoa Dani, Akira Hori, and Gerhard Franz Walter. *Principles of neural aging*. Elsevier, Amsterdam New York, 1997.

[14] Guy Drouin. **Chromatin diminution in the copepod Mesocyclops edax: diminution of tandemly repeated DNA families from somatic cells**. *Genome*, **49**(6):657–665, 2006. Available from: `http://article.pubs.nrc-cnrc.gc.ca/ppv/RPViewDoc?issn=1480-3321{&}volume=49{&}issue=6{&}startPage=657{&}ab=y`.

[15] Donald Freed, Eric L Stevens, and Jonathan Pevsner. **Somatic mosaicism in the human genome**. *Genes*, **5**(4):1064–1094, 2014.

[16] James W. Gaubatz. **Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells**. *Mutation Research DNAging*, **237**(5-6):271–292, 1990.

[17] James W Gaubatz and Sonia C Flores. **Tissue-specific and age-related variations in repetitive sequences of mouse extrachromosomal circular DNAs**. *Mutation Research*, **237**:29–36, 1990.

[18] Bernat Gel and Eduard Serra. **Genome analysis karyoploteR : an R / Bioconductor package to plot customizable genomes displaying arbitrary data**. *Bioinformatics*, **33**(19):3088–3090, 2017.

[19] Géraldine Gentric and Chantal Desdouets. **Polyploidization in liver tissue**. *American Journal of Pathology*, **184**(2):322–331, 2014. Available from: `http://dx.doi.org/10.1016/j.ajpath.2013.06.035`.

[20] C Goday and S Pimpinelli. **The occurrence, role and evolution of chromatin diminution in nematodes**. *Parasitology Today*, **9**(9):319–322, 1993.

[21] Clara Goday and M. Rosario Esteban. **Chromosome elimination in sciarid flies**. *BioEssays*, **23**(3):242–250, 2001.

[22] Anthony Griffiths. *Introduction to genetic analysis*. W.H. Freeman & Company, New York, NY, 2015.

[23] Margaret L Hoang, Isaac Kinde, Cristian Tomasetti, K Wyatt Mcmahon, and Thomas A Rosenquist. **Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing**. pages 3–8, 2016.

[24] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. **Fast mapping of short sequences with mismatches, insertions and deletions using index structures**. *PLoS Computational Biology*, **5**(9):1–10, 2009.

[25] August Yue Huang, Zheng Zhang, Adam Yongxin Ye, Yanmei Dou, Linlin Yan, Xiaoxu Yang, Yuehua Zhang, and Liping Wei. **MosaicHunter: Accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples**. *Nucleic Acids Research*, **45**(10), 2017.

[26] A. John Iafrate, Lars Feuk, Miguel N. Rivera, Marc L. Listewnik, Patricia K. Donahoe, Ying Qi, Stephen W. Scherer, and Charles Lee. **Detection of large-scale variation in the human genome**. *Nature Genetics*, **36**(9):949–951, 2004.

[27] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. **The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community**. *Genome Biology*, **17**(1):1–11, 2016. Available from: `http://dx.doi.org/10.1186/s13059-016-1103-0`.

[28] T James and P Jill. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. **14**(2):178–192, 2012.

[29] Adam F Johnson, Ha T Nguyen, and Reiner A Veitia. **Causes and effects of haploinsufficiency**. *Biological Reviews*, (94):1774–1785, 2019.

[30] D. Karolchik. **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Research*, **32**(90001):493D–496, 2003.

[31] Dhruv Kaushal, James J a Contos, Kai Treuner, Amy H Yang, Marcy a Kingsbury, Stevens K Rehen, Michael J McConnell, Masaru Okabe, Carrolee Barlow, and Jerold Chun. **Alteration of gene expression by chromosome loss in the postnatal mouse brain.** *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **23**(13):5599–606, 2003. Available from: `http://www.ncbi.nlm.nih.gov/pubmed/12843262`.

[32] David G Kent and Anthony R Green. **Order Matters : The order of somatic mutations influences cancer evolution**. *Cold Spring Harb Perspect Med*, **7**(4):1–16, 2017.

[33] Junho Kim, Sangwoo Sanghyeon Kim, Hojung Nam, Sangwoo Sanghyeon Kim, and Doheon Lee. **SoloDel: A probabilistic model for detecting low-frequent somatic deletions from unmatched sequencing data**. *Bioinformatics*, **31**(19):3105–3113, 2015.

[34] M Kloc and B Zagrodzinska. **Chromatin elimination–an oddity or a common mechanism in differentiation and development?** *Differentiation; research in biological diversity*, **68**:84–91, 2001.

[35] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. **VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing**. *Genome Research*, **22**(3):568–576, 2012.

[36] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. **LUMPY: A probabilistic framework for structural variant discovery**. *Genome Biology*, **15**(6):1–19, 2014.

[37] Heng Li and Richard Durbin. **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics*, **26**(5):589–595, 2010.

[38] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*, **25**(16):2078–2079, 2009.

[39] Michael A Lodato, Rachel E Rodin, Craig L Bohrson, Michael E Coulter, Alison R Barton, Minseok Kwon, Maxwell A Sherman, Carl M Vitzthum, Lovelace J Luquette, Chandri N Yandava, Pengwei Yang, Thomas W Chittenden, Nicole E Hatem, Steven C Ryu, Mollie B Woodworth, Peter J Park, and Christopher A Walsh. **Aging and neurodegeneration are associated with increased mutations in single human neurons**. *Science*, **559**(February):555–559, 2018.

[40] Douglas R Lowy, Warren A Kibbe, D Ph, and Louis M Staudt. **Toward a Shared Vision for Cancer Genomic Data Robert**. *The New England journal of medicine*, **375**(12):1109–1112, 2018.

[41] M J McHeyzer-Williams, M. G. McLean, P A Lalor, and G J V Nossa. **Antigen-driven B Cell Differentiation In Vivo**. *Journal of experimental medicine*, **178**(July), 1993.

[42] Heather C Mefford and Barbara J Trask. **The complex structure and dynamic evolution of human subtelomeres**. *Nature Reviews Genetics*, **3**(February):1–12, 2002.

[43] Brandon Milholland, Xiao Dong, Lei Zhang, Xiaoxiao Hao, Yousin Suh, and Jan Vijg. **Differences between germline and somatic mutation rates in humans and mice**. *Nature Communications*, **8**(May):1–8, 2017. Available from: http://dx.doi.org/10.1038/ncomms15183.

[44] Henrik Devitt Møller, Marghoob Mohiyuddin, Iñigo Prada-Luengo, M Reza Sailani, Jens Frey Halling, Peter Plomgaard, Lasse Maretty, Anders Johannes Hansen, Michael P Snyder, Henriette Pilegaard, Hugo Y K Lam, and Birgitte Regenberg. **Circular DNA elements of chromosomal origin are common in healthy human somatic tissue**. *Nature Communications*, **9**(1):1–12, 2018. Available from: http://dx.doi.org/10.1038/s41467-018-03369-8.

[45] Tobias Mourier. **Potential movement of transposable elements through DNA circularization**. *Current Genetics*, **62**(4):697–700, 2016.

[46] Fritz Muller, Vincent Bernard, and Heinz Tobler. **Chromatin diminution in nematodes**. *BioEssays ICSU Press*, **18**(2):133–138, 1996.

[47] Jae Yong Nam, Nayoung K D Kim, Sang Cheol Kim, Je Gun Joung, Ruibin Xi, Semin Lee, Peter J Park, and Woong Yang Park. **Evaluation of somatic copy number estimation tools for whole-exome sequencing data**. *Briefings in Bioinformatics*, **17**(2):185–192, 2016.

[48] NCBI and Bethesda (MD). **Entrez Programming Utilities Help [Internet]**. *National Center for Biotechnology Information*, (Md):1–161, 2010. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25501/.

[49] Maeve O'Huallachain, Konrad J Karczewski, Sherman M Weissman, Alexander Eckehart Urban, and Michael P Snyder. **Extensive genetic variation in somatic human tissues**. *Proceedings of the National Academy of Sciences*, **109**(44):18018–18023, 2012. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23043118{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3497787http://www.pnas.org/cgi/doi/10.1073/pnas.1213736109.

[50] William R Pearson and David J Lipmant. **Improved tools for biological sequence comparison**. *Proceedings of the National Academy of Sciences of the United States of America*, **85**(April):2444–2448, 1988.

[51] Aaron R Quinlan and Neil Kindlon. **bedtools: a powerful toolset for genome arithmetic**. *University of Utah*, 2008. Available from: `https://bedtools.readthedocs.io/`.

[52] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. **DELLY: Structural variant discovery by integrated paired-end and split-read analysis**. *Bioinformatics*, **28**(18):333–339, 2012.

[53] José Reyes, Laura Gómez-romero, Ximena Ibarra-soria, Kim Palacios-flores, Luis R Arriola, and Alejandro Wences. **Context-dependent individualization of nucleotides and virtual genomic hybridization allow the precise location of human SNPs**. *PNAS*, **108**(37):15294–15299, 2011.

[54] Natalie Saini and Dmitry A Gordenin. **Somatic Mutation Load and Spectra : A Record of DNA Damage and Repair in Healthy Human Cells**. **686**(June):672–686, 2018.

[55] Xinwei She, Julie E Horvath, Zhaoshi Jiang, Ge Liu, Terrence S Furey, Laurie Christ, Royden Clark, Tina Graves, Cassy L Gulden, Can Alkan, Jeff A Bailey, Cenk Sahinalp, Mariano Rocchi, David Haussler, Richard K Wilson, Webb Miller, Stuart Schwartz, and Evan E Eichler. **The structure and evolution of centromeric transition regions within the human genome**. *Nature*, **430**(August):857–864, 2004.

[56] A F A Smit, R Hubley, and P Green. **RepeatMasker**, 2018. Available from: `http://repeatmasker.org`.

[57] T F Smith and M S Watrman. **Identification of Common Molecular Subsequences**. *Journal Molecular Biology*, (147):195–197, 1981.

[58] Todd J Treangen and Steven L Salzberg. **Repetitive DNA and next-generation sequencing: computational challenges and solutions**. *Nat Rev Genet.*, **13**(1):36–46, 2013.

[59] Jose M C Tubio. **Somatic structural variation and cancer**. *Briefings in Functional Genomics*, **14**(5):339–351, 2015.

[60] Reiner A Veitia, Diddahally R Govindaraju, Samuel Bottani, and James A Birchler. **Aging : Somatic Mutations , Epigenetic Drift and Gene Dosage Imbalance**. *Trends in Cell Biology*, **27**(4):299–310, 2017. Available from: `http://dx.doi.org/10.1016/j.tcb.2016.11.006`.

[61] L Wolpert. *Principles of development*. Oxford University Press, Oxford, United Kingdom New York, NY, United States of America, 2015.

[62] Kimie Yamazaki Yamamoto, Kiyoko Yamazaki and Yoshihiro Kato. **Changes of chromosomes during the early neural development of a japanese newt , Cynops Pyrrhogaster**. *Development, Growth and Differentiation*, **22**(2):79–92, 1980.

[63] Zhen Zhang, Jianxin Wang, Junwei Luo, Xiaojun Ding, Jiancheng Zhong, Jun Wang, Fang Xiang Wu, and Yi Pan. **Sprites: Detection of deletions from sequencing data by re-aligning split reads**. *Bioinformatics*, **32**(12):1788–1796, 2016.